

Moral judgements on human and autonomous drivers' decisions in unavoidable collisions scenarios

Augusto Bovesi
a.bovesi@student.unisi.it
DISPOC - Università di Siena
Siena, Italy

Alice Calabretto
a.calabretto@student.unisi.it
DISPOC - Università di Siena
Siena, Italy

Alice Stroppa
a.stroppa@student.unisi.it
DISPOC - Università di Siena
Siena, Italy

Gian Maria Adamo
g.adamo8@student.unisi.it
DISPOC - Università di Siena
Siena, Italy

Filippo Canepa
f.canepa@student.unisi.it
DISPOC - Università di Siena
Siena, Italy

Stefano Guidi
stefano.guidi@unisi.it
DISPOC - Università di Siena
Siena, Italy

ABSTRACT

Moral and ethical issues are constantly arising in assessing the perception of autonomous vehicles and their “behaviour” in daily traffic situations. A debated question is how individuals perceive the choices taken by autonomous vehicles (AVs) in life-threatening scenarios. In an online experiment ($N = 232$) we tested whether the actions taken by an AV or a human driver in realistic road-accident scenarios are judged according to different standards. In addition, multiple factors were manipulated, such as the number of pedestrians crossing the road, the number of occupants inside the vehicle and the outcome of the choice.

The results highlight a preference for human agents with respect to AVs. In addition, there is a significant difference in the type of agent, with respect to the utilitarian principle. The human self-sacrifice attitude is appreciated to a different degree, according to the type of individuals saved (pedestrians or occupants), but is confirmed to be a powerful factor in moral evaluations. The results might have implications for increasing acceptability of AVs.

CCS CONCEPTS

• **Social and professional topics;** • **Applied computing** → **Transportation;**

KEYWORDS

autonomous vehicles; behavioural decisions; ethical dilemma; utilitarianism; road-accident scenarios

ACM Reference Format:

Augusto Bovesi, Alice Calabretto, Alice Stroppa, Gian Maria Adamo, Filippo Canepa, and Stefano Guidi. 2024. Moral judgements on human and autonomous drivers' decisions in unavoidable collisions scenarios. In *European Conference on Cognitive Ergonomics (ECCE 2024)*, October 08–11, 2024, Paris, France. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3673805.3673827>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ECCE 2024, October 08–11, 2024, Paris, France

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1824-3/24/10

<https://doi.org/10.1145/3673805.3673827>

1 INTRODUCTION

Autonomous Vehicles (AVs) are an emerging transport method whose adoption could bring many benefits, like reducing road accidents caused by fatigue, drink-driving and human error [6, 7, 15]. Nevertheless, crashes remain unavoidable independently of the driver's nature. Car accidents caused by AVs need an adequate moral framework, also because research indicates that individuals may be hesitant to invest in AV technology unless there exists a transparent moral framework guiding the decision-making processes of autonomous driving systems [23]. Hence, the challenge lies in devising morally acceptable decisions aligning with public expectations [1, 23]. In fact, there exists a behavioural inconsistency of a preference to buy a passenger-protective car over a prosocial one or to pay a premium to purchase a “selfish” vehicle [14], despite judging the latter as more moral [3].

Most studies on humans and AV behaviours in road-accident dilemmas are shaped according to the Trolley Dilemma [8, 25, 26], a dilemmatic problem designed to investigate the utilitarian and deontological preference on whether to sacrifice one person to save a larger number of lives. In a recent study [21], participants were asked to judge how morally adequate the actions of an AV and a human driver were in a series of road-accident scenarios. The action chosen by the agent and the number of pedestrians on the road were the variables considered. The results highlighted that the human drivers' actions were judged as morally superior to those of an AV, and that both actions were judged in a positive moral way whenever they were in line with the utilitarian principles [13, 21]. However, other studies [9, 17] suggest the use of double standards in evaluations of human and AV decisions, with the former being judged with respect to deontological principles and the latter more with respect to utilitarian principles.

The endorsement of the utilitarian principle can be also influenced by the involvement of self-sacrifice [24] of the decision maker. A number of studies [4, 12, 14], considering also the role of perspective taking [19], showed that participants tend to endorse the utilitarian choice less when it requires self-sacrifice. On the other hand, self-sacrifice seems to be preferred when the number of spared lives increases [20], when the spared individuals are women and children, and in time-pressure situations [27].

Further investigations seem therefore needed to understand both the presence of double standards and the role of self-sacrifice.

2 THE STUDY

This study aimed to replicate and extend recent findings [21] regarding the moral evaluation of decisions taken by human drivers and AVs in critical traffic situations presenting moral dilemmas. To investigate the recourse to the utilitarian principle in moral judgments and the role of self-sacrifice we manipulated the choice, the number of pedestrians on the road, and the number of occupants in the car.

2.1 Research Questions and Hypotheses

The study was designed and conducted to answer the following questions.

- RQ1 *Is there a bias for decisions by human drivers over decisions by AVs?* Based on the results of previous studies [17], the decisions of humans should be judged more favourably than the corresponding decisions by AVs, either due to a general preference for human decisions in these types of dilemmas, or due to a general aversion toward machines [16] (HP1).
- RQ2 *Are decisions by AVs evaluated more according to a utilitarian principle than human decisions?* If the decision of an AV is evaluated according to utilitarian principles more than decisions of a human driver [17], the effect of the number of pedestrians on the road should be stronger for an AV than for a human driver, and the effect of the number of occupants should be stronger for a human driver than for an AV (HP2).
- RQ3 *Is self-sacrifice positively considered in the moral evaluations of human drivers' decisions?* Participants should rate more positively the decision of the human driver to give their life to spare the life of one or more pedestrians, based on [20, 21, 24] (HP3).

3 METHOD

The experiment was conducted using an online questionnaire implemented in English with Google Forms. Overall, the completion of the forms took about 6 minutes. The experiment was conducted according to the Declaration of Helsinki and approved by the Ethical Committee of the University of Siena (*actn.05/2024*). Informed consent was obtained from each participant before the experiment.

3.1 Participants

We recruited 232 participants for the experiment. 165 participants were initially recruited through snowballing via messaging apps and social media platform. 68 additional participants were recruited using Prolific.com and paid (0.75£) for their participation in the study. The additional sample had the same gender distribution of the initial sample, but the mean and standard deviation of age were lower ($M = 30.2, SD = 7.8$) than in the initial sample ($M = 35.6, SD = 14.4$). The age of participants ranged from 18 to 99 ($M = 34.0, SD = 13.0$). Of those who chose to disclose their gender ($N = 221, 95.3\%$), 46.6% identified with the feminine gender, 52.9% with the masculine gender, and one participant as non-binary. The sample included participants from 41 different countries ($NA = 7.3\%$), with the most represented ones being the US (15.4%), the UK (14.9%), and Italy (13%).

3.2 Design and Procedure

Participants were randomly assigned to rate the moral acceptability of the actions of either a human driver ($N = 114$) or an AV ($N = 118$), in 8 inevitably fatal collision scenarios. The scenarios resulted from the combination of the levels of three factors: the type of action chosen (killing the occupants of the car or killing pedestrians on the road), the number of pedestrians on the road (1 or 2) and the number of human occupants in the vehicle (1 or 2). The experiment had thus a 4-way $2 \times 2 \times 2 \times 2$ mixed-design (agent \times choice \times number of pedestrians \times number of occupants). To control order and sequence effects the order of the scenarios was counterbalanced across participants using an 8×8 balanced Latin square. Participants initially read an introductory text, which stated that human drivers or autonomous vehicles have to handle traffic situations, including the accidents. In the autonomous-vehicle conditions, AV was defined as “a self-driving car capable of driving completely by itself and moving through traffic without the need for human intervention”. Participants were then told they would have to evaluate several scenarios with a similar structure. The instructions for the human-driver condition were stated as shown in Figure 1A. In the autonomous-vehicle condition, the instructions were formulated in the same way, but “person” was replaced by “autonomous vehicle”. In each scenario, the agent - and possibly a passenger - drove on a single-lane road and was suddenly confronted with an obstacle and at least one pedestrian on the road. The possible actions were sacrificing the person/s inside the vehicle to save the pedestrian/s by crashing into the obstacle, or sacrificing the pedestrian/s to save the person/s inside the vehicle. In the trials, the scenarios were depicted as abstract sketches from a bird's eye view (Figure 1B) presented along a text vignette (Figure 1C). There were either one or two pedestrians and/or one or more occupants. The yellow arrow represented the available action taken by the agent. The red skull indicated who was intended to be sacrificed. The images were adapted from [21]. Below each image and the text vignette, participants were asked to evaluate the action of the agent from a moral perspective, in same way as in [21]. The question listed the agent, the action and the action's consequences. For example, “How do you evaluate, from a moral point of view, the action of the driver to save him/herself and sacrifice the person on the street?”. Participants were asked to complete the sentence “From a moral point of view, I perceive the action as...” by choosing a rating on a Likert scale ranging from 1 to 6, with 1 being “very reprehensible” and 6 being “very justifiable”. The same operation was repeated for each scenario in the assigned module until the eighth and last scenario. The next and final section asked for general demographic information, joint with the due reminder of the anonymity of the data gathering process: age, gender, occupational status and nationality.

4 RESULTS

The data were analysed using a mixed, factorial, $2 \times 2 \times 2 \times 2$ ANOVA, using the multivariate approach and including agent (human driver or autonomous vehicle) as a between-subject factor, and the following within-subject factors: choice (sacrifice the pedestrian/s or sacrifice the person/s inside the car), number of pedestrians on the road (one or two), and number of occupants in the car (one or two). All the analyses were conducted using R (v. 4.0.2). The results

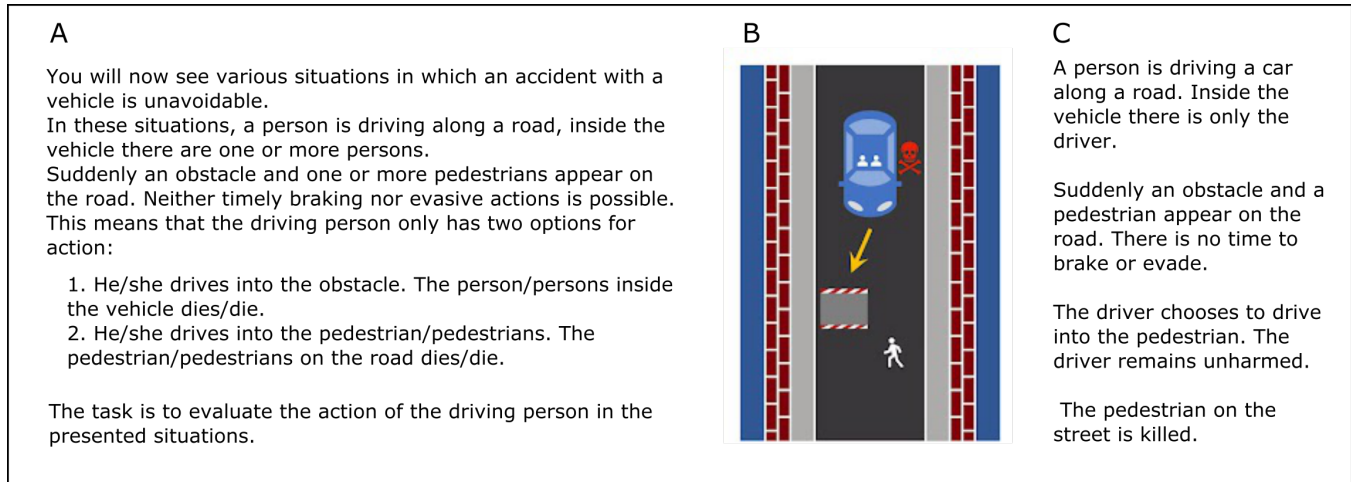


Figure 1: A) The instructions presented to participants with the structure of the scenarios in the human driver condition. B) Pictorial and C) textual representations of an example scenario used in the experiment.

of the ANOVA showed significant main effects of agent ($F_{1,222} = 10.15, p = .002, \eta_p^2 = .044$, small), choice ($F_{1,222} = 54.13, p < .001, \eta_p^2 = .196$ large) and number of occupants ($F_{1,222} = 20.89, p < .001, 2p = .086$ medium). The results also showed significant two-way interactions between agent and choice ($F_{1,222} = 10.25, p = .002, \eta_p^2 = .044$, small), between choice and number of occupants ($F_{1,222} = 202.91, p < .001, \eta_p^2 = .478$, large), and between choice and number of pedestrians ($F_{1,222} = 83.39, p < .001, \eta_p^2 = .273$, large). Lastly, the 3-way interactions between agent, choice and number of pedestrians ($F_{1,222} = 6.94, p = .009, \eta_p^2 = .030$, small), and between agent, choice and number of occupants ($F_{1,222} = 10.29, p = .002, \eta_p^2 = .044$, small) were also significant, and so was the 4 way interaction ($F_{1,222} = 4.59, p = .033, \eta_p^2 = .020$, small). In the next sections we unpack the nature of these effects.

4.1 Effect of agent and choice

The actions of the human driver ($M = 3.65, SE = 0.06$) were evaluated as more morally justifiable than the actions of the autonomous vehicle ($M = 3.38, SE = 0.06, t(222) = -3.19, p = 0.002$). Sacrificing the person inside the vehicle was evaluated more favourably ($M = 3.98, SE = 0.07$) than sacrificing the pedestrian/s ($M = 3.05, SE = 0.08, t(222) = 7.36, p = 0$). Post-hoc comparisons following the significant interaction between agent and choice, however, showed that sacrificing the driver was evaluated more favourably when the agent was a human driver ($M = 4.32, SE = 0.11$) than when it was an AV ($M = 3.64, SE = 0.1, t(222) = -4.57, p = 0$), while sacrificing the pedestrian/s was evaluated similarly whether the action was made by a human driver ($M = 2.98, SE = 0.11$) or by an AV ($M = 3.12, SE = 0.11, t(222) = 0.88, p = 0.379$).

4.2 Effects of number of occupants

The actions were evaluated as more morally justifiable when there was only one human occupant in the car ($M = 3.6, SE = 0.05$) than when there were two ($M = 3.43, SE = 0.04, t(222) = 4.57, p = 0$).

However, the analysis of significant interaction between choice and number of occupants showed that an increase in the number of occupants in the car led to a significant increase in the moral evaluation of sacrificing the pedestrian/s (for one occupant: $M = 2.79, SE = 0.08$; for two occupants: $M = 3.31, SE = 0.08, t(222) = -9.81, p = 0$) and to a significant decrease in the moral evaluation of sacrificing the persons in the car (for one occupant: $M = 4.41, SE = 0.08$; for two occupants: $M = 3.55, SE = 0.08, t(222) = 12.99, p = 0$). A 2-way interaction contrast revealed that the effect of the number of occupants was significantly larger for the choice of sacrificing the person/s in the car than for the one of killing the pedestrian/s ($t(222) = 14.25, p = 0$).

The analysis of the interaction between agent and number of occupants showed instead that only for a human driver the action was evaluated more favourably when there was a single occupant in the car ($M = 3.76, SE = 0.07$) than when there were two occupants ($M = 3.54, SE = 0.06, t(222) = 4.32, p = 0, d = 0.58$). These evaluations were also significantly different when the agent was an AV ($t(222) = 2.09, p = 0.038, d = 0.28$), though the effect size was smaller.

4.3 Effects of number of pedestrians

The main effect of the number of pedestrians was not significant ($F_{1,222} = 0.17, p = .679, \eta_p^2 < .001$), but the analysis of the simple effects of number of pedestrians for the different choices showed that this was due to the significant crossed-over interaction between choice and number of pedestrians. An increase in the number of pedestrians on the road, in fact, led to a significant decrease in the moral evaluation of sacrificing the pedestrian/s (for one pedestrian: $M = 3.25, SE = 0.08$; for two pedestrians: $M = 2.85, SE = 0.08, t(222) = 6.85, p = 0$) and to a significant increase in the moral evaluation of sacrificing the persons in the car for one pedestrian: $M = 3.76, SE = 0.08$; for two pedestrians: $M = 4.2, SE = 0.08, t(222) = -8.6, p = 0$).

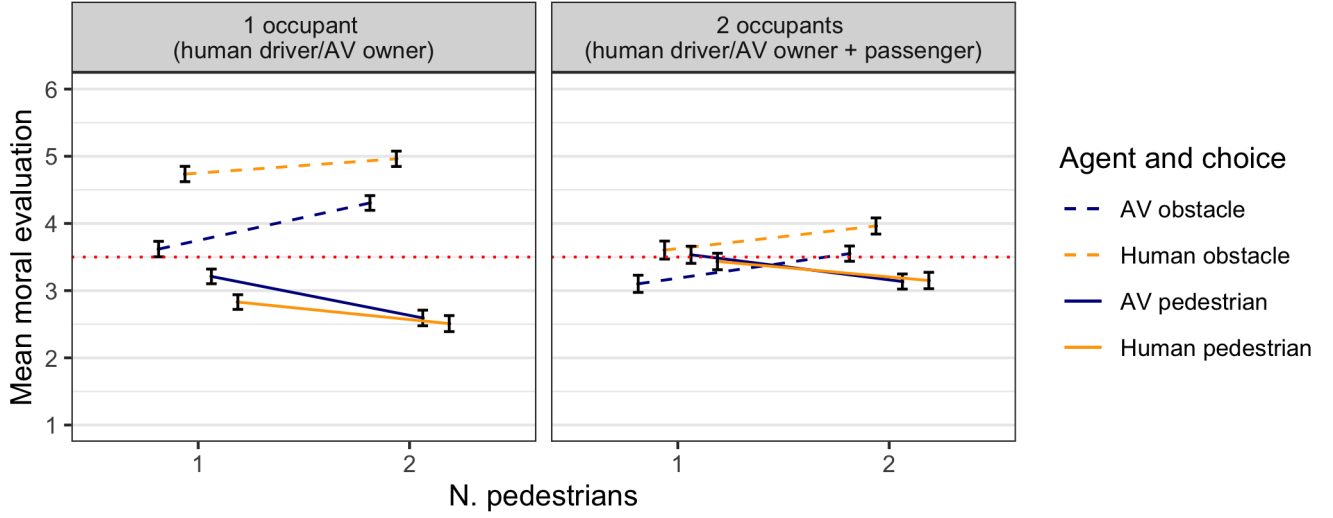


Figure 2: Plots of the mean moral evaluations of the decisions of sacrificing the pedestrian or sacrificing the person inside the car, as a function of decision maker, number of occupants in the car, and number of pedestrians on the road. The moral-evaluation scale ranged from “very reprehensible” (1) to “very justifiable” (6). Error bars represent 95% within-subjects confidence intervals for the means.

4.4 Differences in the moral evaluations of decisions by actor

Furthermore, to analyse the two significant three-way interactions between *agent*, *choice* and *number of pedestrians* and between *agent*, *choice* and *number of occupants*, an ANOVA was conducted between 2 (choice) \times 2 (number of pedestrians) \times 2 (number of occupants) for each of the two agents separately. The results showed significant main effects of *choice* (AV: $F_{1,117} = 9.96, p = .002, \eta_p^2 = .078$, HD: $F_{1,105} = 48.38, p < .001, \eta_p^2 = .315$) for both agents and *number of occupants* (AV: $F_{1,117} = 4.21, p = .042, \eta_p^2 = .035$, HD: $F_{1,105} = 19.42, p < .001, \eta_p^2 = .156$); and large significant interactions between *choice* and *number of pedestrians* (AV: $F_{1,117} = 50.80, p < .001, \eta_p^2 = .303$, HD: $F_{1,105} = 39.29, p < .001, \eta_p^2 = .272$) on the one hand, and between *choice* and *number of occupants* on the other hand (AV: $F_{1,117} = 67.57, p < .001, \eta_p^2 = .366$, HD: $F_{1,105} = 137.26, p < .001, \eta_p^2 = .567$). For the autonomous vehicle, there was also a significant 3-way interaction between *choice*, *number of pedestrians* and *number of occupants* ($F_{1,117} = 4.68, p = .032, \eta_p^2 = .038$). In Figure 2 are plotted the mean evaluations of the choice in all the experimental conditions.

The effects of *choice* and *number of occupants* were always larger for the human driver than for the autonomous vehicle. Consistently with [21], the effect of the *number of pedestrians* on the moral evaluation of the action of sacrificing the person/s inside the car was significant and positive for both agents. It was more pronounced though, for AVs ($t(117) = -7.01, p = 0, d = 0.65$) than for human drivers ($t(105) = -5.34, p = 0, d = 0.52$). While the corresponding effects on the evaluations of the choice of killing the pedestrian were both significant and negative, but similar in size (AV: $d = 0.48$, HD: $d = 0.47$).

Also, the effect of the *number of occupants* was significant for both agents. In this case, the effect was stronger for the evaluation of the actions of the human driver (sacrificing the person/s in the car: $t(105) = 10.72, p = 0, d = 1.05$, killing the pedestrian/s: $t(105) = -8.35, p = 0, d = 0.81$) than for those of the AV (sacrificing the person/s in the car: $t(117) = 7.36, p = 0, d = 0.68$, sacrificing the pedestrian/s: $t(117) = -5.63, p = 0, d = 0.52$). Moreover, a significant 3-way interaction contrast ($t(222) = 3.21, p = .002$) showed that the comparison by choice of the pairwise contrast for the effect of the number of occupants was different for a human driver than for an AV, and greater for the human driver. In other words, the asymmetry in the strength of effect of the *number of occupants* was more pronounced for a human driver than for an AV.

To analyse the significant 4-way interaction we tested simple contrasts on the *agent* factor at all the combinations of the levels of *choice*, *number of occupants* and *number of pedestrians*. The results showed that the moral bias was only present for the choice of sacrificing the person inside the car ($p < .05 - .001$). For the alternative choice, in the majority of the conditions, no significant bias was found ($p \geq .596 - .932$). With one pedestrian and one occupant the bias was in the opposite direction, with the action of the AV ($M = 3.21, SE = 0.12$) being evaluated as slightly but significantly more morally justifiable than the action of the human driver ($M = 3.21, SE = 0.12, t(222) = 2.17, p = 0.031$).

5 DISCUSSION

The aim of this study was first of all (RQ1) to verify the moral evaluation bias that favours decisions made by human agents over analogous decisions by autonomous vehicles in road situations in which an accident is unavoidable, presenting a moral dilemma [21].

The results of our study only partially confirm this bias (HP1). This effect, in fact, was only reliably significant for the choice of sacrificing the person/s in the car, and even for this choice it was generally small, except with one occupant in the car. For the alternative choice of sacrificing the pedestrian in most cases the bias was not significant, and even when it was, when there was one pedestrian on the road and one occupant in the car, the bias was again small and in the opposite direction: the autonomous vehicle was evaluated more favourably than the human driver. This study also aimed at testing differences in the moral evaluations of decisions by humans and autonomous vehicles in relationship to the recourse of utilitarian principles (RQ 2). Previous studies yielded conflicting results on this topic, with some studies finding asymmetries in the importance of utilitarian principles for humans and machines [9, 17], and others failing to consistently find significant and meaningful differences even with larger samples [13, 21]. Our results seem more in line with the former group. First of all, results show that the effect of the number of pedestrians in the choice of sacrificing the person in the car was stronger for the AV than for the human driver, as it would be expected if AV were evaluated more according to a utilitarian principle. The effect of the number of occupants in the car, instead, was stronger for the human driver than for the AV, for both choices. It is interesting to note that when two occupants were in the car, the utilitarian choice of sacrificing a single pedestrian on the road (sparing 1 life with respect to the other choice) for a human driver was evaluated not significantly more morally justifiable than the choice of sparing the pedestrian and self-sacrificing killing also the other passenger in the car. This leads us to the third research question in our study (RQ 3), aimed at investigating the effect of self-sacrifice in evaluations of human decisions. Self-sacrificing one's life to spare someone else's one should be favourably considered in the moral evaluations (HP 3). Indeed [21] proposed that for human decision makers a positive appraisal of self-sacrifice could increase the moral evaluation of that action even in the case of one pedestrian (and no utilitarian advantage of the choice over killing the pedestrian), causing the effect of the number of pedestrians (and of lives spared by the choice) to be smaller for humans' decisions than for AVs' ones. This explanation can be indeed applied to our results for the scenarios in which only the driver was in the car. We observe an effect of self-sacrifice also when it involves killing the passenger, when there is no utilitarian advantage across the alternative choices (in each case two individuals are killed, and two are spared), and possibly also when it violates the utilitarian principle, causing more victims. When there is only one pedestrian and two occupants in the car, in fact, the action of sacrificing the persons in the car was evaluated significantly more positively when made by a human driver (self-sacrificing) than when an AV made it, although in this case the effect is confounded with the positive bias toward human decisions. Indeed, we found that for a human driver, killing one pedestrian on the road to save their own life and the life of the passenger, despite the utilitarian advantage, was not judged significantly more morally justifiable than the action of self-sacrificing and killing the other occupant to spare the life of the pedestrian on the road. It might be that in this condition, the positive evaluation of self-sacrifice compensates for the fact that the action resulted in a higher number of victims than the (selfish) action of killing the pedestrian. Also, participants evaluated the AV

killing the person(s) inside the car slightly, but significantly, more favourably than the AV killing the pedestrians, when no utilitarian advantage was present in either of the choices. It thus seems that self-sacrifice is valued also for autonomous vehicles, consistently with what was found by [4, 12, 14, 20, 24].

Overall, it seems that a tendency to evaluate more positively decisions that include self-sacrifice might be enough to almost fully explain our results. If there was a general bias for human decisions, we should have found it also for the choice of sacrificing the pedestrians, and regardless of utilitarian considerations, while we only find this bias for the decision involving self-sacrifice. Unfortunately, in our experiment it is not possible to separate an effect of self-sacrifice from a general bias for human decisions. The fact that the decision of a human driver to kill a pedestrian to self-preserve their life was evaluated as less morally justifiable than the corresponding decision by an AV, seems to suggest that pursuing self-preservation in this context, instead of being seen as a valid (or at least understandable) moral justification for the action of killing someone [20], might be seen as egoistic behaviour, and accordingly devalued from a moral point of view. From the results of our study, however, we cannot exclude the hypothesis that double standards are used in evaluating humans and machines [17, 18, 24], the latter giving greater weight to utilitarian principles, in favour of the hypothesis that a positive bias for self-sacrifice, coupled with a negative bias for egoistic actions is sufficient to account for the participants' responses [24]. Lastly, the significant interactions between choice and number of occupants and between agent, choice and number of occupants showed that for both agents the effect of the number of occupants was larger for the action of killing the persons inside the car than for the action of killing the pedestrians. This asymmetry was stronger in the evaluations of the actions of the human driver than in the evaluations of the actions of the AV, maybe due to the perception of a possible relationship between the car driver and the passenger [19]. In fact, it can be that from a moral point of view the choice of sacrificing a passenger the human driver knows is heavier than to swerve and kill a presumably unknown pedestrian. In our study, however, we did not control for the type of relationship between the driver and the passenger, therefore further investigations should be required to confirm this hypothesis.

5.1 Limitations

Our study has some limitations that need to be acknowledged.

First of all, in our experiment it was not possible to separate an effect of self-sacrifice from a general bias for human decisions. Future research should thus try to clarify the relative importance of self-sacrifice and of a general bias toward human over machine decisions, separating and contrasting these factors by design, and including measures of individual attitudes toward and trust in technology, and qualitative judgements over the choices. Secondly, our sample is not culturally homogeneous, and while this should increase the generalisability of our findings, literature [2] shows that there are cultural differences in moral judgements of autonomous vehicles' decisions. Moreover, our sample was obtained using two different recruiting methods, although previous studies [5] suggest this should not matter much for the results. Further limitations are intrinsic to moral dilemmas, which despite being extensively used

in research have not been exempt from criticism [11, 22]. On the one hand, the scenarios depicted are abstractions, often devoid of potentially relevant contextual information or factors (e.g. age, gender ...), and in which outcomes are presented as certainties. On the other hand, research has shown that moral decisions are based also on intuitive judgements and emotions that might precede proper rational moral reasoning [10], which would be used only post-hoc to justify decisions.

6 CONCLUSIONS

Our results suggest that judgements about critical decisions in road accident scenarios made by human drivers are different from judgements about the same decisions made by an AV. Concerning the reason for this difference, however, our study reveals a quite complex picture. A bias for human decisions is not always present, and sometimes the opposite bias can be found, possibly due to a bias against egoism. Moreover, different standards might be used in evaluating the actions of human drivers and AVs, with the former being judged more on the basis of utilitarian considerations than the latter. In recent years there have been increasing efforts to develop and deploy on the road fully autonomous vehicles for personal or commercial use. Finding ways to reduce biases in the moral evaluation of AV decisions in critical road-situations presenting moral dilemmas could be important for the ultimate success of these efforts. As rare as these situations might be, it is in fact likely that accidents will receive large media coverage, especially in the early phases of adoption. If previous research has suggested increasing anthropomorphism as a means to reduce a possible negative bias against autonomous vehicles [21], our findings suggest another possible strategy. If self-sacrifice is considered in the moral evaluations, in fact, maybe increasing the perceived consciousness of the AV, might allow it to benefit from the bias. Although it is unclear whether self-sacrifice of a conscious machine could be positively valued from a moral point of view, future research should at least try to test this hypothesis empirically.

REFERENCES

- [1] Nadia Adnan, Shahrina Md Nordin, Mohamad Ariff bin Bahrudin, and Murad Ali. 2018. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. *Transportation Research Part A: Policy and Practice* 118 (Dec. 2018), 819–836. <https://doi.org/10.1016/j.tra.2018.10.019>
- [2] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (Nov. 2018), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>
- [3] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2016. The social dilemma of autonomous vehicles. *Science* 352, 6293 (June 2016), 1573–1576. <https://doi.org/10.1126/science.aaf2654>
- [4] Giovanni Bruno, Andrea Spoto, Lorella Lotto, Nicola Cellini, Simone Cutini, and Michela Sarlo. 2023. Framing self-sacrifice in the investigation of moral judgment and moral emotions in human and autonomous driving dilemmas. *Motivation and Emotion* 47, 5 (Oct. 2023), 781–794. <https://doi.org/10.1007/s11031-023-10024-3>
- [5] Benjamin D. Douglas, Patrick J. Ewell, and Markus Brauer. 2023. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLOS ONE* 18, 3 (03 2023), 1–17. <https://doi.org/10.1371/journal.pone.0279720>
- [6] Veljko Dubljević. 2020. Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles. *Science and Engineering Ethics* 26, 5 (Oct. 2020), 2461–2472. <https://doi.org/10.1007/s11948-020-00242-0>
- [7] Daniel J. Fagnant and Kara Kockelman. 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy and Practice* 77 (2015), 167–181. <https://doi.org/10.1016/j.tra.2015.04.003>
- [8] Philippa Foot. 2002. *Virtues and Vices*. Oxford University Press. <https://doi.org/10.1093/0199252866.001.0001>
- [9] Stefano Guidi, Enrica Marchigiani, Sergio Roncato, and Oronzo Parlangeli. 2021. Human beings and robots: are there any differences in the attribution of punishments for the same crimes? *Behaviour & Information Technology* 40, 5 (April 2021), 445–453. <https://doi.org/10.1080/0144929X.2021.1905879>
- [10] Jonathan Haidt. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108, 4 (2001), 814–834. <https://doi.org/10.1037/0033-295X.108.4.814>
- [11] Johannes Himmelreich. 2018. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice* 21, 3 (June 2018), 669–684. <https://doi.org/10.1007/s10677-018-9896-4>
- [12] Uijong Ju and Sanghyeon Kim. 2024. Willingness to take responsibility: Self-sacrifice versus sacrificing others in takeover decisions during autonomous driving. *Heliyon* 10, 9 (May 2024), e29616. <https://doi.org/10.1016/j.heliyon.2024.e29616>
- [13] Jamy Li, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram F. Malle. 2016. From Trolley to Autonomous Vehicle: Perceptions of Responsibility and Moral Norms in Traffic Accidents with Self-Driving Cars. 2016–01–0164. <https://doi.org/10.4271/2016-01-0164>
- [14] Peng Liu and Jinting Liu. 2021. Selfish or Utilitarian Automated Vehicles? Deontological Evaluation and Public Acceptance. *International Journal of Human-Computer Interaction* 37, 13 (Aug. 2021), 1231–1242. <https://doi.org/10.1080/10447318.2021.1876357>
- [15] Chiara Lucifora, Giorgio Mario Grasso, Pietro Perconti, and Alessio Plebe. 2020. Moral dilemmas in self-driving cars. *Rivista internazionale di Filosofia e Psicologia* Vol. 11, n. 2 (Aug. 2020), 238–250. <https://doi.org/10.4453/ripf.2020.0015>
- [16] Hasan Mahmud, A. K. M. Najmul Islam, Syed Ishtiaque Ahmed, and Kari Smolander. 2022. What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change* 175 (Feb. 2022), 121390. <https://doi.org/10.1016/j.techfore.2021.121390>
- [17] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, Portland Oregon USA, 117–124. <https://doi.org/10.1145/2696454.2696458>
- [18] Bertram F. Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. 2016. Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 125–132. <https://doi.org/10.1109/HRI.2016.7451743>
- [19] Rose Martin, Petko Kusev, and Paul Van Schaik. 2021. Autonomous vehicles: How perspective-taking accessibility alters moral judgments and consumer purchasing behavior. *Cognition* 212 (July 2021), 104666. <https://doi.org/10.1016/j.cognition.2021.104666>
- [20] Maike M. Mayer, Raoul Bell, and Axel Buchner. 2021. Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLOS ONE* 16, 12 (Dec. 2021), e0261673. <https://doi.org/10.1371/journal.pone.0261673>
- [21] Maike M. Mayer, Axel Buchner, and Raoul Bell. 2023. Humans, machines, and double standards? The moral evaluation of the actions of autonomous vehicles, anthropomorphized autonomous vehicles, and human drivers in road-accident dilemmas. *Frontiers in Psychology* 13 (Jan. 2023), 1052729. <https://doi.org/10.3389/fpsyg.2022.1052729>
- [22] Sven Nyholm and Jilles Smids. 2016. The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice* 19, 5 (Nov. 2016), 1275–1289. <https://doi.org/10.1007/s10677-016-9745-2>
- [23] Jonathan Robinson, Joseph Smyth, Roger Woodman, and Valentina Donzella. 2022. Ethical considerations and moral implications of autonomous vehicles and unavoidable collisions. *Theoretical Issues in Ergonomics Science* 23, 4 (July 2022), 435–452. <https://doi.org/10.1080/1463922X.2021.1978013>
- [24] Sonya Sachdeva, Rumen Iliev, Hamed Ekhtiari, and Morteza Dehghani. 2015. The Role of Self-Sacrifice in Moral Dilemmas. *PLOS ONE* 10, 6 (June 2015), e0127409. <https://doi.org/10.1371/journal.pone.0127409>
- [25] Judith Jarvis Thomson. 1985. The Trolley Problem. *The Yale Law Journal* 94, 6 (May 1985), 1395. <https://doi.org/10.2307/796133>
- [26] Judith Jarvis Thomson and The Hegeler Institute. 1976. Killing, Letting Die, and the Trolley Problem. *Monist* 59, 2 (1976), 204–217. <https://doi.org/10.5840/monist197659224>
- [27] Huarong Wang, Dongqian Li, Zhenhang Wang, Jian Song, Zhan Gao, and David C. Schwebel. 2023. Who Should We Choose to Sacrifice, Self or Pedestrian? Evaluating Moral Decision-Making in Virtual Reality. In *Engineering Psychology and Cognitive Ergonomics*, Don Harris and Wen-Chin Li (Eds.). Vol. 14018. Springer Nature Switzerland, Cham, 560–572. https://doi.org/10.1007/978-3-031-35389-5_39 Series Title: Lecture Notes in Computer Science.