Should it stay, or swerve? Trading off lives in dilemma situations involving autonomous cars

Habla, Wolfgang; Kataria, Mitesh; Martinsson, Peter; Roeder, Kerstin

Published in:
Health Economics (United Kingdom)

Link to article, DOI:
10.1002/hec.4802

Publication date:
2024

Document Version
Publisher's PDF, also known as Version of record

Received: 25 February 2023 | Revised: 22 August 2023 | Accepted: 17 December 2023

DOI: 10.1002/hec.4802

RESEARCH ARTICLE

Health Economics WILEY

Should it stay, or swerve? Trading off lives in dilemma situations involving autonomous cars

Wolfgang Habla[1] | **Mitesh Kataria**[2] | **Peter Martinsson**[2,3] | **Kerstin Roeder**[4]

[1]Baden-Wuerttemberg Cooperative State University (DHBW), Villingen-Schwenningen, Germany

[2]University of Gothenburg, Gothenburg, Sweden

[3]Technical University of Denmark, Kongens Lyngby, Denmark

[4]University of Augsburg, Augsburg, Germany

Correspondence
Wolfgang Habla.
Email: wolfgang.habla@dhbw-vs.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: 139943784; Stiftelsen för Miljöstrategisk Forskning

Abstract

Using a representative survey with 1317 individuals and 12,815 moral decisions, we elicit Swedish citizens' preferences on how algorithms for self-driving cars should be programmed in cases of unavoidable harm to humans. Participants' choices in different dilemma situations (treatments) show that, at the margin, the average respondent values the lives of passengers and pedestrians equally when both groups are homogeneous and no group is to blame for the dilemma. In comparison, the respondent values the lives of passengers more when the pedestrians violate a social norm, and less when the pedestrians are children. Furthermore, we explain why the average respondent in the control treatment needs to be compensated with two to six passengers spared in order to sacrifice the first pedestrian, even though she values the lives of passengers and pedestrians equally at the margin. We conclude that respondents' choices are highly contextual and consider the age of the persons involved and whether these persons have complied with social norms.

KEYWORDS

choice experiments, ethical preferences, random utility model, relative values of life, robot cars, self-driving cars

1 | INTRODUCTION

Every day, medical staff make decisions on who should be treated and which resources are to be used. These decisions imply that some individuals will get priority over others, that is, some will be treated but others not, or some will get better treatment than others. To help medical staff make these decisions, many countries have developed guidelines for priority settings. The guidelines are often general and hence the final decision is affected by the judgment by the medical staff on duty. In a new era where robots and artificial intelligence are slowly replacing human decisions in the healthcare sector, these decisions can no longer be left to subjective judgment. Instead, exact algorithms with predefined decision rules are needed. From a health policy perspective, self-driving cars are the type of robot that will most likely have the largest impact on health, and hence these cars need clear rules on how to decide in terms of priority setting. For example, if a pedestrian is crossing the road just in front of a self-driving car, should the self-driving car

swerve and injure or kill the people in the car, or not swerve and injure or kill the pedestrian? The objective of this paper is to empirically investigate people's preferences for how these self-driving cars should be programmed, using a representative survey conducted in Sweden.

The key issue is what decision a self-driving car, or any type of robot, should make when the choice is only between bad health outcomes. In the iconic thought experiment known as the "trolley dilemma" (e.g., Engisch, 1930; Foot, 1967), a trolley is heading toward a group of five people that are stuck on a railway track. Without any intervention, the trolley will overrun and kill these five people. However, there is a lever that redirects the trolley to a sidetrack where only one person is standing. This person will be killed if the lever is pulled. The ethical question then is whether or not the lever should be pulled to save the five and sacrifice the one, or vice versa. This dilemma situation and modified versions of it have been extensively investigated in the literature (see, e.g., Navarrete et al., 2011; Swann et al., 2010; Thomson, 1985).

For a long time, the trolley dilemma was criticized for being too distant from real-life moral situations to be useful for policymakers or educational purposes. However, given that the arrival of fully autonomous vehicles is imminent, situations similar to the trolley dilemma have almost become a reality. And once self-driving cars are fully available, that is, deployed on public roads, situations may arise in which harm to one or several humans is unavoidable—thus, the vehicle needs to be programmed to make a split-second decision whether to harm one or more pedestrians, or one or more passengers. While moral philosophy and law can guide society as to how algorithms for self-driving cars *should* be programmed, it is a matter of public acceptance which ethical rules *will* eventually be implemented.[1] Such ethical rules include treating all people alike and saving the larger number of people, or discriminating against some individuals based on, for example, their age, or their compliance with social norms or laws.

Self-driving cars have an enormous potential to save lives, since motor vehicle accidents "represent the eight leading cause of death globally" (WHO, 2018, p. ix) and even the leading cause of death for children and young adults aged 5–29 years. Overall, more than 1.35 million people die each year due to accidents, and up to 50 million are injured. Most of these deaths and injuries could be avoided, and self-driving cars can considerably contribute to lowering morbidity and mortality rates from accidents, as they are able to make decisions much faster and in a more informed way than any human. Against this background, self-driving vehicles could "create one of the most important public health advances of the 21st century" (Fleetwood, 2017, p. 532).[2] While dilemma situations certainly represent only a minor share in the large death toll in traffic crashes, self-driving cars are in principle able to reduce the number of casualties in such situations if programmed this way. What is more, the widespread adoption of these cars hinges on their acceptability by the public, which is severely affected by the discussions surrounding the ethical trade-offs in these dilemma situations. Furthermore, we can expect that more and more decisions concerning our health and safety will be delegated to algorithms and artificial intelligence, for example, when it comes to surgeries carried out by robots. The acceptance of such technologies that could be highly beneficial to society as a whole will also hinge on how they are programmed and trained, particularly with respect to ethical decisions. Whenever resources for these technologies are scarce or dilemma situations arise, questions will arise about whether individual attributes, such as age, or the violation of certain norms and laws, should be considered for the execution of a particular treatment.

In this paper, we use a choice experiment to explore the ethical preferences of a representative sample of the Swedish population. The ethical preferences are revealed by asking our subjects how they would like self-driving cars to be programmed for dilemma situations. In our choice experiment, respondents were confronted with a scenario in which a fully autonomous car with one or more passengers on board is driving down a road at the speed limit. One or more pedestrians are on a pedestrian crossing but they are already on the other lane—thus, the car does not need to slow down in order to pass by them. Suddenly and unexpectedly, a huge rock falls from a neighboring slope, directly ahead of the car. The stopping distance is too long for the car to avoid a crash, so the car is therefore faced with the following dilemma: either it continues forward and crashes into the rock, which implies certain death for the passenger (s); or it swerves to the left, in which case it kills all pedestrians on the road. In any case, harm to humans is unavoidable.[3]

After being presented with this scenario, respondents were asked to choose—for varying numbers of people in the car and on the road—whether they would prefer the self-driving car to be programmed such that it continues forward, or swerves to the left (swerving to the right was not an option, as we will explain later). More precisely, we presented 10 choice sets to each survey participant. In each choice set, there were up to five passengers and up to five pedestrians. We then asked participants to state how the self-driving car should be programmed in each choice set. The above scenario served as the control scenario (termed CONTROL treatment in the following) in which only adults of the same age were involved. Notably, neither pedestrians nor passengers could be blamed for causing this dilemma situation, as they were

all "rightfully" on the road. In two further treatments, we introduced blame and different ages. In the BLAME treatment, the scenario was modified in such a way that the pedestrians were jaywalking, thus violating a social norm. As in the CONTROL treatment, all passengers and pedestrians were adults of the same age. In the AGE treatment, the passengers of the car were again adults but the pedestrians were children who were *not* jaywalking, so no responsibility for the dilemma situation could be assigned to either group. Survey participants were randomly put into three groups, each of which received one of the above treatments.

Before we discuss the results, it is instructive to distinguish between two different moral principles: consequentialism, and, more specifically, utilitarianism on the one hand, and deontology on the other hand. According to consequentialism, "morality is all about producing the right kinds of overall consequences" (Internet Encyclopedia of Philosophy, 2019). Utilitarianism is one example of consequentialism in the sense that actions are judged as moral to the degree that they produce the best overall outcomes (utility) across all parties (Mill, 1861/1998). In line with earlier studies on dilemma situations of autonomous vehicles (e.g., Bonnefon et al., 2016), we consider a certain choice to be utilitarian if it saves the highest number of lives, and we consider someone to be utilitarian if she consistently makes utilitarian choices.[4] In contrast to this, deontological ethics holds that the morality of an action should be based on whether an action itself is right or wrong under a series of rules, rather than be based on the outcome of the action. A traditional trolley dilemma can more readily demonstrate what a deontological rule is. Consider, for example, the case where a man could be killed to harvest his organs in order to save the lives of five others. In this case, the deontological rule could be that you should never kill this man, because he just happens to be at the same place as the five patients. A utilitarian, by contrast, would think differently.

Descriptively, we find that a large majority of choices in all treatments were in favor of sacrificing the passenger(s) rather than the pedestrian(s). In the CONTROL treatment, 78% of the choices involved sacrificing the passenger(s). In the BLAME treatment, in which the pedestrians were responsible for causing the dilemma situation, a much smaller share of 61% of all choices favored sacrificing the passenger(s). By contrast, when the pedestrians were children, 89% of the choices spared the pedestrians. Furthermore, we simultaneously find a majority making utilitarian choices, which indicates that there may be conflicting preferences in the sample population, which cannot be disentangled based on descriptive statistics. This impression is amplified when considering that the majority of all utilitarian choices actually spare pedestrians rather than passengers, and the majority of all "Continue forward" choices (which spare the pedestrians) are utilitarian. Finally, only 20% of all respondents are utilitarians, in the sense that they always chose the alternative that saved the maximum number of lives in the choice sets where this choice was possible.

To understand and disentangle the preferences of the respondents based on their choices, we estimate a discrete choice model and calculate the marginal rate of substitution between passenger and pedestrian deaths. We find an average marginal rate of substitution close to unity in the CONTROL treatment, that is, the average respondent in our sample is willing to sacrifice one passenger in order to save one (more) pedestrian. This suggests that people, on average and at the margin, value the lives of pedestrians and passengers equally. However, this only holds for this very specific scenario in which the groups of passengers and pedestrians are homogeneous. When pedestrians are responsible for the potential collision with the car (BLAME treatment), the average marginal rate of substitution is smaller than unity, indicating that the lives of pedestrians are valued less than those of passengers. When children cross the street and are not to blame for the accident (AGE treatment), people tend to assign higher priority to them as compared to the adult passenger(s), resulting in a marginal rate of substitution substantially above unity.

In addition to marginal costs, there also is a discrete cost of sacrificing pedestrians, where the cost is expressed in terms of how many passengers need to be saved in order to sacrifice the first pedestrian present in the scenario. When we take this into account, we find that, depending on the treatment, a number of two to six passengers needs to be saved in order for the first pedestrian to be sacrificed. In the CONTROL treatment, for example, the average respondent prefers the car to continue forward when there is one pedestrian on the road, unless the car seats about four passengers or more. This means that with three passengers, the car will continue forward and save the pedestrian. Adding a fourth passenger, however, changes the preference to "swerve", while adding a second pedestrian would change the decision back to "continue forward". In the BLAME treatment, by contrast, the average respondent is already indifferent between sacrificing the passengers and the one pedestrian on the road when there are two passengers in the car. For more than two, the respondent would sacrifice the pedestrian. In the AGE treatment, the average respondent requires six passengers to sacrifice the one child on the road, but a second child on the road would reverse this decision. Our results thus suggest that there is a certain threshold for the difference between the number of passengers and pedestrians—a threshold beyond which people are more utilitarian and are more likely to value lives equally. This is in line with the idea of so-called "threshold deontology" (Zamir & Medina, 2010).

Overall, most of our results indicate that, on average, respondents do not apply utilitarian ethics when making choices. Only in the CONTROL treatment do respondents, at the margin, value the lives of passengers and pedestrians equally, which is in line with utilitarianism. In the other two treatments, however, our results reveal that ethical preferences are stronger toward the lives of children and weaker toward the lives of people that are themselves to blame for their unpleasant situation.[5] The huge discrete cost of sacrificing the first pedestrian in the scenario in all treatments also speaks against applying utilitarian ethics, unless for choice sets in which the number of people in one group significantly exceeds the number of people in the other group.

The rest of the paper is structured as follows. In the next section, we review the related literature. Section 3 introduces the sampling framework and contains a detailed description of the survey. Section 4 presents descriptive statistics on the choices people made in the survey. In Section 5, we present logistic regressions in order to analyze what variables best predict utilitarian choices and the likelihood of consistently making utilitarian choices. Section 6 considers the marginal and discrete cost of sacrificing one pedestrian (in terms of how many passengers need to be saved). In Section 7, we discuss the results and potential caveats of this study. Section 8 provides the conclusion.

2 | RELATED LITERATURE

Our study is related to two strands of the literature. The first one is the literature on priority setting in the healthcare sector, and in particular people's view on priority settings. Second, our paper links to the literature on traditional trolley dilemmas.

Priority setting in the healthcare sector rests on different ethical principles (e.g., Williams & Cookson, 2000). The most important question is whether priority setting should be based only on efficiency in terms of saving people's lives and avoiding or reducing injuries, or whether personal characteristics and responsibility should also affect priority setting. For example, the Swedish healthcare system is defined in the Health Care Act from 1982 and the objective is to provide "good health and care on equal terms for the entire population". Priority setting was described by a governmental-commissioned report (Einhorn et al., 1995) and ratified by the government in 1997. According to this report, priority setting should be guided by the principle of equal value, the principle of need, and the principle of efficiency, in the order they are stated. Similar types of guidelines for priority setting also exist in other countries. In practical terms, this rules out that personal characteristics or responsibility should matter for priority setting. However, people's views on priority setting typically differ from the above guidelines. For example, Cropper et al. (1994) ask people from the general public to choose between different projects targeting certain accidents or diseases, which would result in different numbers of saved people of different ages. The authors find that people place more weight on saving younger persons, but that this relationship is hump-shaped. More specifically, for saving a 30-year-old, the median respondent was willing to sacrifice more 60-year-olds than for saving a 20- or 40-year-old. Similar findings have been obtained by studies that examine attitudes toward health interventions at different ages (see, e.g., Baltussen et al., 2006; Carlsson et al., 2010; Dolan & Tsuchiya, 2005; Johansson-Stenman et al., 2011). More closely related to our research question is the study by Johansson-Stenman and Martinsson (2008) who present different life-saving road investment projects to survey participants. Their results not only show that the lives of younger individuals are consistently given higher values than the lives of older individuals, but also that the lives of pedestrians are valued more highly than the lives of drivers of the same age. Covey et al. (2010) examine how the values people place on preventing fatalities from rail accidents are affected by the extent to which victims are responsible for their death. The authors show that when people are to blame for the accident, their value of life is (about 50%) lower. This, however, is not the case when the person acting irresponsibly is a child. In our study, respondents also value pedestrians' lives less when they are to blame for the dilemma situation to arise, and more when the pedestrians are children. Similarly, for example, Olsen et al. (2003), Cappelen and Norheim (2005), Dolan and Tsuchiya (2009), Edlin et al. (2012) and Gu et al. (2015) discuss the role of personal characteristics and responsibility for priority setting and present empirical results indicating that people seem to generally agree that these factors should matter.

The literature on trolley experiments, as the one described in the introduction (or modifications of it, such as 'The fat man', 'The fat villain' or 'Transplant'), has more recently also been framed in the context of autonomous cars. Generally, people confronted with such moral dilemmas have been found to be utilitarian in the sense that they would like to maximize the number of lives saved. The difference between traditional trolley dilemmas and those that involve an autonomous car is that in the case of the latter, the ethical decision is delegated to an algorithm long before a dilemma situation actually occurs.

In the context of autonomous cars, Bonnefon et al. (2016) find that people approve of autonomous vehicles that sacrifice their passengers in case a higher number of pedestrians can be saved, but they would not buy vehicles that are programmed like that. This finding points to a social dilemma behind the ethical dilemma: self-driving vehicles may be fully beneficial to society only if they are programmed to save the highest number of lives, but this is not accepted by the majority of potential car owners. While Bonnefon et al. asked participants about trade-offs involving mostly one versus one or one versus ten (or two vs. 20) people who could either be saved or sacrificed,[6] we asked participants to make choices involving different (and more realistic) numbers of people (one, two, three, four, five). Like Bonnefon et al., we find that most choices are utilitarian. However, the vast majority of respondents do not make utilitarian choices *throughout* their choice sets (where this is possible). Using a discrete choice model, we uncover the preferences of the respondents and show that respondents are, on average, *not* utilitarian, unless under very specific circumstances.

Frank et al. (2019) examine decision-making biases that could arise in moral dilemmas of autonomous vehicles. Similar to Bonnefon et al. (2016), they find that people's personal perspectives, that is, whether they imagine themselves to be bystanders or involved in the dilemma, affect their decisions. Another main finding of Frank et al. (2019) is that situational factors, such as the presence of children or the lawfulness of pedestrians' behavior, play an important role in people's moral preferences. In particular, if children are present, decisions are biased in their favor. Social norm violations by a pedestrian, as indicated by walking out in front of a car or jaywalking at a red light, are found to increase the likelihood of this pedestrian being sacrificed. In the present paper, we also examine situations in which children are present and pedestrians violate traffic norms and confirm that situational factors are an important determinant of people's moral choices.

In a study based on data collected from the Moral Machine website, Awad et al. (2018) elicit cultural and individual differences with respect to the choices made in trolley dilemmas involving self-driving cars. On the Moral Machine website, visitors are confronted with several randomly assigned scenarios that are similar to the ones in Bonnefon et al. (2016) or the ones in our paper, but those also include choices between killing people and killing animals, or killing people that might be more or less useful for society (such as medical doctors or criminals). Awad et al. find that moral preferences with respect to the presented dilemma situations do indeed differ across countries and cultures. In particular, they identify three major clusters of countries that exhibit significant ethical variation (the Western, Eastern, and Southern clusters). These clusters differ in their preferences to spare younger characters rather than older characters and higher-status characters rather than lower-status characters, but they share the same (though weak) preference for sparing pedestrians over passengers, or the lawful over the unlawful. In contrast to cultural differences, Awad et al. find little evidence that individual variations, such as age, education, gender, income, political or religious views, have sizable impacts on the choices made in the Moral Machine experiment. While the results of Awad et al. might suffer from self-selection bias, we have a representative sample of the Swedish population.

3 | SAMPLING FRAMEWORK AND SURVEY DETAILS

In this section, we describe the panel which we used for the survey, the sample of respondents and the survey questions and choice sets that the participants were given.

3.1 | Panel

We made use of the Citizen Panel (henceforth CP; in Swedish: Medborgarpanelen), which is an online panel survey run by the Laboratory of Opinion Research (LORE) at the Faculty of Social Sciences at the University of Gothenburg. Each survey carried out by LORE consists of several specific studies as well as a number of more general questions that are not included in a specific study. Specific studies are either survey experiments using random assignment or panel studies that span over several waves of the CP. The CP has approximately 55,000 active respondents all over Sweden. Our survey was included in the 24th CP, which ran between March 21 and April 17, 2017 (for a detailed description see Martinsson et al., 2017). Overall, this wave contained five studies (including ours) and 26 additional general questions. A reminder to participate in the survey was sent twice during that period.

3.2 | Survey details

The survey was conducted in Swedish (for a complete transcription into English see the Supporting Information S1) and contained two introductory pages on which potential advantages of self-driving cars were described (better use of road capacity, collision avoidance, availability for people who cannot drive themselves). Furthermore, respondents were informed about how these cars collect information on the surroundings. They were also informed that the sensors of the car could distinguish between adults and children.[7] Also, a picture of a self-driving car by Google, and a picture of how the surroundings would be seen by the car through radar were displayed. We then emphasized that while traffic will, on average, become safer with these cars and take fewer lives, accidents can never be completely eliminated.

On the next page of the survey, one of three alternative setups was randomly assigned to participants. We will refer to these setups as treatments. In all treatments, a self-driving car with one or more passengers on board is heading down a road at the speed limit, and one or more pedestrians are crossing the road when—suddenly and unexpectedly—a huge rock falls from the neighboring slope onto the direct path of the car. In this situation, the self-driving car has only two options, as the stopping distance is too long to avoid the collision. It can either **continue forward** and head straight into the rock, which implies certain death for the passenger(s), or it can **swerve** to the left and hit the pedestrian(s), which implies certain death for the pedestrian(s).[8]

The three treatments are depicted in Figure 1. In the CONTROL treatment (left panel), all passengers and pedestrians are adults of the same age (which we did not specify), and the pedestrian(s) "rightfully" cross the road on a pedestrian crossing. In the BLAME treatment (panel in the middle), the pedestrian crossing is regulated by a traffic light, and this traffic light shows red. Thus, the pedestrian(s) commit(s) a social norm violation, with the social norm being that jaywalking is not allowed. Finally, in the AGE treatment, the pedestrians are 10-year-old children, and there is no traffic light at the pedestrian crossing. Passengers and pedestrians in all treatments are homogeneous within a group with respect to age, that is, all passengers or pedestrians are the same age (we did not specify their gender and graphically illustrated them in a gender-neutral way).[9] In the pictorial representations of the three setups, we ensured and emphasized that all adults or all children are the same height and wear the same clothes.

Next, in each treatment group, respondents were randomly shown one of three tables with 10 choice sets each. The choices to be made involved different numbers of pedestrians and passengers, for example, three passengers and four pedestrians. The minimum number of people in each group (pedestrians and passengers) was one, the maximum number was five. Respondents were then asked for each of the 10 choice sets to decide how the self-driving car should be programmed, that is, whether it should save the lives of the pedestrians or those of the passengers. Figure 2 exemplarily displays the different tables for the CONTROL treatment (the tables were the same across treatments, with only the two last columns having different headers, depending on the specific treatment).

The choice sets were arranged using a full factorial design such that the number of pedestrians and the number of passengers are uncorrelated in each alternative over all choice sets. Moreover, all possible combinations between the numbers of passengers and pedestrians for each of the alternatives were allowed. We stressed in the survey that

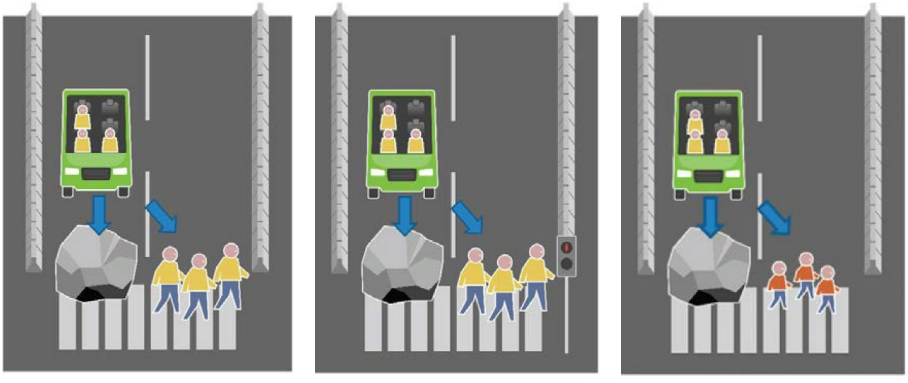


FIGURE 1 The three different setups (treatments) randomly displayed to respondents.

Table 1	Table 2	Table 3	Car continues forward and all adult passengers die	Car swerves to the left and all adult pedestrians die
Scenario 1: 2 passengers meet 1 pedestrian	Scenario 1: 5 passengers meet 1 pedestrian	Scenario 1: 2 passengers meet 3 pedestrians	❍	❍
Scenario 2: 5 passengers meet 2 pedestrians	Scenario 2: 2 passengers meet 2 pedestrians	Scenario 2: 4 passengers meet 2 pedestrians	❍	❍
Scenario 3: 3 passengers meet 3 pedestrians	Scenario 3: 1 passenger meet 3 pedestrians	Scenario 3: 3 passengers meet 1 pedestrian	❍	❍
Scenario 4: 1 passenger meets 4 pedestrians	Scenario 4: 2 passengers meet 4 pedestrians	Scenario 4: 5 passengers meet 4 pedestrians	❍	❍
Scenario 5:4 passengers meet 5 pedestrians	Scenario 5: 3 passengers meet 5 pedestrians	Scenario 5: 5 passengers meet 5 pedestrians	❍	❍
Scenario 6: 1 passenger meets 1 pedestrian	Scenario 6: 1 passenger meets 4 pedestrians	Scenario 6: 1 passenger meets 1 pedestrian	❍	❍
Scenario 7: 1 passenger meets 2 pedestrians	Scenario 7: 1 passenger meets 2 pedestrians	Scenario 7: 3 passengers meet 2 pedestrians	❍	❍
Scenario 8: 5 passengers meet 3 pedestrians	Scenario 8: 4 passengers meet 3 pedestrians	Scenario 8: 1 passenger meets 3 pedestrians	❍	❍
Scenario 9: 3 passengers meet 4 pedestrians	Scenario 9: 4 passengers meet 1 pedestrian	Scenario 9: 4 passengers meet 4 pedestrians	❍	❍
Scenario 10: 2 passengers meet 5 pedestrians	Scenario 10: 1 passenger meets 5 pedestrians	Scenario 10: 1 passenger meets 5 pedestrians	❍	❍

FIGURE 2 The three tables randomly displayed to different groups of respondents for the CONTROL treatment (only one table was shown to each respondent, so either Table 1, 2, or 3).

respondents should not imagine themselves being involved in the scenarios (as passengers of the car or pedestrians) but rather take on the role of an impartial spectator. If the respondents did not answer all choice sets, they were prompted to do so before proceeding to the next page of the survey (but could also proceed without making a choice). Hence, the respondent could opt out of difficult choices by simply ignoring one or several choice sets (see footnote 10).

3.3 | Sample

For our study, 2500 panel members received an invitation to take part in the survey, of whom 1574 responded and 1458 finished the survey. 1317 respondents filled in at least one choice set. They constitute our sample for the analysis. The full sample had been stratified on sex, age (18–75 years), and education, and the survey was programmed in Qualtrics.

In Table 1, some descriptive statistics on the sample are presented. A share of 54% of the sample respondents were males. Respondents were, on average, 49.42 years old, and 23% of them attended the university for three years or more. Furthermore, 68% of our respondents had children, and they earned an average personal monthly income before taxes of SEK 31,088 (approximately USD 3200). 19% of the respondents were retired, 8% were students and 2% were unemployed. We also include several variables that we collected ourselves following the survey questions, such as car ownership or mode choice for commuting. 87% of our sample had a driver's license, and more than two thirds owned a car. Furthermore, 51% commuted to work by car, while 13% commuted by foot. 36% had heard about dilemma situations as depicted in our choice experiment before they took part in the survey. Furthermore, our sample seems to be fairly representative of the Swedish population, as indicated in the last column of Table 1, but also with respect to variables other than those used for stratification.

We next examine if our sample is balanced across treatments. For the different treatment groups (CONTROL, BLAME, AGE), Table 2 presents descriptive statistics similar to Table 1. The means differ only slightly across treatments, and statistical tests reveal that the sample is balanced for all variables, not only for those that were used for the stratification (see Table A1 in the Appendix).

4 | DESCRIPTIVE STATISTICS OF THE RESULTS

In this section, we present descriptive statistics on the choices made by respondents, and on the shares of respondents who consistently made certain choices throughout their choice sets. In particular, we are interested in how many choices are utilitarian or favor either of the two groups of people present in the treatments (passengers or pedestrians).

TABLE 1 Descriptive statistics and comparison with the Swedish population (where available).

Variable	Description	Full sample				Swedish population mean
		Mean	sd	Min	Max	
Male	= 1 if male	0.54	0.50	0	1	0.50[c]
Age	Age in years	49.42	15.28	18	75	49.19[d]
University education[a]	= 1 if at university for 3 years or more	0.23	0.42	0	1	0.18[d]
Personal monthly income[b]	Gross, in SEK	31,088	17,226	0	95,000	25,203[e]
Children	= 1 if yes	0.68	0.47	0	1	
Unemployed	= 1 if unemployed	0.02	0.14	0	1	0.07[f]
Retired	= 1 if retired	0.19	0.39	0	1	0.20[f]
Student	= 1 if student	0.08	0.27	0	1	0.08[g]
Additional covariates (survey questions):						
Driver's license	= 1 if driver's license	0.87	0.34	0	1	0.87[g]
Car ownership	= 1 if owning a car	0.69	0.46	0	1	
Commute to work by car	= 1 if commute by car	0.51	0.50	0	1	0.57[g]
Commute to work by foot	= 1 if commute by foot	0.13	0.34	0	1	0.09[g]
Heard about dilemma	= 1 if heard before	0.36	0.48	0	1	
Observations (# of individuals)		1317				

[a]The sample was stratified on several branches of the Swedish education system, not only on university education as indicated by this variable. We constructed a binary variable for later purposes.

[b]Income was given in intervals of varying size. The mean here is computed based on the assumption that income is uniformly distributed within each interval, so we took the midpoint of each interval. This could also be the reason why the mean income in the sample deviates from the Swedish population mean.

[c]Taken from the website of Statistics Sweden for 2016. For the whole Swedish population, that is, including age cohorts 1–17 and 76 and older.

[d]Own computations for all people aged 18 or older, based on population statistics from the website of Statistics Sweden for 2016.

[e]Taken from the website of Statistics Sweden for 2015.

[f]Own computations based on the population aged 65 years and older relative to the total population for 2015 (taken from the United Nations World Population Prospects: The 2017 Revision).

[g]Own computations based on data from the Swedish National Travel Survey (RVU Sweden) 2011–2016.

4.1 | Continue forward, swerve, and utilitarian choices

In total, the respondents made 12,815 choices.[10] Panel A of Table 3 and Figure 3a illustrate these choices for each treatment group. A share of 78% of all choices in the CONTROL treatment were "Continue forward" and thus entailed certain death for the passenger(s). This share is considerably lower in the BLAME treatment (61%) in which pedestrians were assigned the responsibility for the dilemma situation, and considerably higher in the AGE treatment (89%) in which children were crossing the street on a pedestrian crossing. The differences across treatments are statistically significant at the 1% level.

By contrast, there is little variation across treatments when it comes to the share of utilitarian choices that, by definition, save the lives of the larger group. In the CONTROL treatment, the share of utilitarian choices is 67%, and we observe only slightly lower shares in the BLAME and AGE treatments (66% and 64%, respectively). The differences are statistically significant at the 1% level between CONTROL and AGE treatment and the 5% level between BLAME and AGE treatment, but the difference between CONTROL and BLAME treatment is insignificant. Obviously, whenever there were equal numbers of passengers and pedestrians in a choice set, people could not make a utilitarian choice. Therefore, these choice sets are excluded from the above calculations.

We can also analyze whether respondents made more or less utilitarian choices when the difference between the number of passengers and the number of pedestrians was larger.[11] The largest difference in the numbers between these two groups of people was four (five persons in one group and one in the other group). As one might have anticipated, we can see from Figure 3b and, in more detail, from Table A2 in the Appendix, that respondents made more utilitarian

TABLE 2 Characteristics of treatment groups.

	CONTROL		BLAME		AGE	
	Mean	**sd**	**Mean**	**sd**	**Mean**	**sd**
Male	0.53	0.50	0.54	0.50	0.56	0.50
Age	49.34	15.31	49.84	15.13	49.07	15.42
University education	0.23	0.42	0.22	0.42	0.26	0.44
Personal monthly income[a]	31,812	17,322	30,318	17,261	31,127	17,097
Children	0.67	0.47	0.68	0.47	0.69	0.46
Unemployed	0.02	0.12	0.02	0.15	0.02	0.14
Retired	0.18	0.38	0.20	0.40	0.19	0.40
Student	0.07	0.25	0.08	0.27	0.09	0.28
Additional covariates:						
Driver's license	0.87	0.34	0.88	0.33	0.87	0.34
Car ownership	0.70	0.46	0.68	0.47	0.69	0.46
Commute to work by car	0.49	0.50	0.51	0.50	0.52	0.50
Commute to work by foot	0.12	0.32	0.14	0.35	0.14	0.34
Heard about dilemma before	0.37	0.48	0.37	0.48	0.33	0.47
Observations (# of individuals)	445		441		431	

Note: The mean here is computed based on the assumption that income is uniformly distributed within each interval, so we took the midpoint of each interval.

[a]Income was given in intervals of varying size.

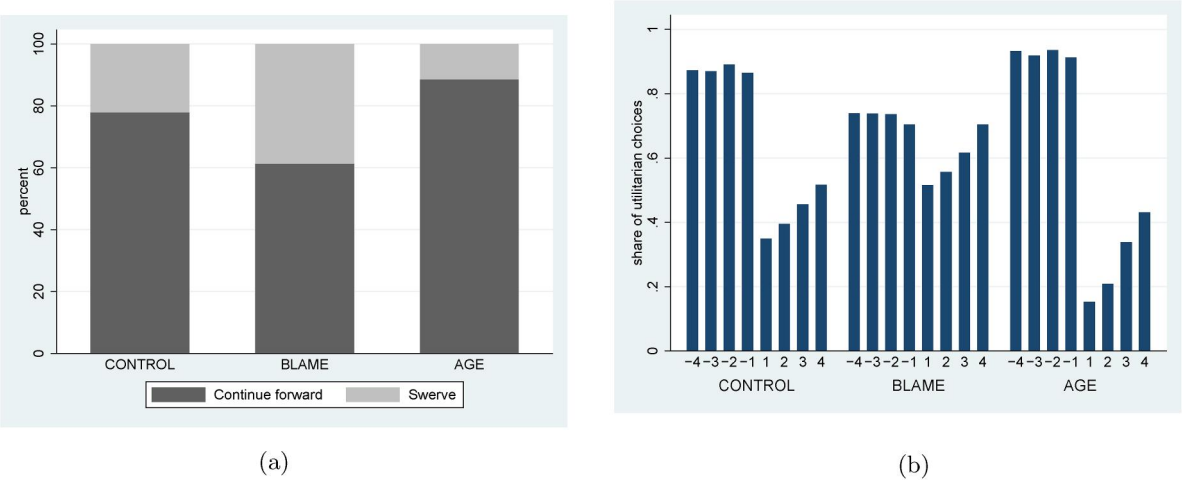
TABLE 3 Shares of choices and shares of respondents with certain choices, across treatments.

	CONTROL	BLAME	AGE
Panel A: Shares of choices			
"Continue forward" (= kill passengers) of those: utilitarian[a]	0.78	0.61	0.89
	0.67	0.71	0.64
"Swerve" (= kill pedestrians) of those: Utilitarian	0.22	0.39	0.11
	0.74	0.62	0.81
Observations[b] (# of choices)	4342	4325	4148
Utilitarian choices[a] of those:	0.67	0.66	0.64
"Continue forward"	0.75	0.64	0.85
"Swerve"	0.25	0.36	0.15
Observations (# of choices)	3545	3543	3444
Panel B: Share of respondents who …			
Always Made utilitarian choices (where possible)[c]	0.238	0.234	0.128
Always chose "continue forward" (= kill passengers)	0.528	0.361	0.661
Always chose "swerve" (= kill pedestrians)	0.083	0.220	0.028
Observations (# of individuals)	445	441	431

[a]This is the share of choices that were utilitarian where this was possible. In the choice sets with equal numbers of passengers and pedestrians to be sacrificed, there was obviously no possibility to make a utilitarian choice. These choices are excluded. The remaining share of choices thus refers to choices that sacrifice the less numerous of the two groups (be they pedestrians or passengers).

[b]Missing choices are excluded here. These are 355 in the full sample.

[c]All respondents are included who made utilitarian choices where this was possible, even if they did not answer all choice sets.



(a) (b)

FIGURE 3 Shares of choices. (a) "Continue forward" (= kill passengers) and "Swerve" (= kill pedestrians) choices. (b) Utilitarian choices and difference in number of people that can be saved (=# of passengers - # of pedestrians in the same choice set).

choices when this difference was larger, particularly for positive differences, that is, when the number of passengers exceeded the number of pedestrians in a choice set. What is more, we see that the share of utilitarian choices is much (and statistically significantly) higher (and quite stable) whenever we have the opposite case, that is, pedestrians were in the majority. This suggests that respondents felt more comfortable sparing the lives of the larger group when this group consisted of pedestrians rather than passengers.

4.2 | People with consistent choices

As each respondent received 10 choice sets, we can determine the share of people who made consistent choices throughout their choice sets. Panel B of Table 3 displays the share of respondents who always made utilitarian choices where this was possible (i.e., either seven, eight, or nine times, depending on which table with choice sets the respondents were assigned to), who always chose to kill the passenger(s) or kill the pedestrian(s). Although a large majority (approximately two thirds) of choices were utilitarian, this was not driven by respondents who consistently made utilitarian choices, since this share only amounts to between 13% and 24%. We also observe that there is basically no difference between the CONTROL and BLAME treatment with respect to the share of utilitarians (23.8 vs. 23.4%, *p*-value of test for significant difference: 0.607), but this share drops to 13% when children were involved in the scenarios (with the difference between the AGE treatment and the other two treatments being statistically significant at the 1% level).

Strikingly, more than 50% of the respondents in the CONTROL treatment always chose to sacrifice the passengers in order to spare the lives of the pedestrians. There are large and significant differences across treatments (see the second row of Panel B in Table 3): in the AGE treatment, this share was even higher (66%), while it was much lower in the BLAME treatment (36%). Thus, respondents seem to have punished social norm violations by pedestrians as indicated by jaywalking in the BLAME treatment, as one might have expected. The presence of children (as pedestrians in the AGE treatment) caused participants to spare this group much more often than in the CONTROL treatment, which implies that people indeed attached different weights to individuals of different age. In contrast to this, the share of people always choosing "Swerve" was considerably lower. This share was 22% in the BLAME treatment and statistically significantly lower in the CONTROL and AGE treatments (8% vs. 3%).

4.3 | Summary

We observed that the majority of choices spare the lives of pedestrians rather than those of passengers, so involved "Continue forward" much more often than "Swerve". This also applies to choices at the individual level where the share of respondents who always chose "Continue forward" is substantially higher than the share of those who always chose

"Swerve". One might conclude from this that the respondents made such choices because they have a preference for sparing the lives of pedestrians. However, these choices could also have been made for the most part because of a preference for saving the maximum number of lives. To see this, consider that among the "Continue forward" choices, a substantial share coincides with the utilitarian choice. Likewise, a large share of choices was utilitarian, and of those choices, the majority of choices coincide with the "Continue forward" choice. While this pattern might to some extent reflect a preference for sparing the lives of the pedestrians, that is, an "aversion to swerve", we conclude from this section that it is not straightforward to disentangle respondents' preferences behind their choices based on the descriptive statistics. Furthermore, the results demonstrate that respondents' moral choices are highly contextual. They depend on the treatment as well as the difference in the numbers of passengers and pedestrians whose lives can be saved.

5 | PREDICTING UTILITARIAN CHOICES

In this section, we analyze what variables best predict utilitarian choices and whether someone is utilitarian, by estimating odds ratios from logit regressions. An odds ratio greater than one increases the likelihood of making a utilitarian choice or being utilitarian, while an odds ratio below one decreases this likelihood.

In more detail, we estimate the following equation for the probability that individual i makes a utilitarian choice:

$$\text{logit}(P(X_i = 1)) = \alpha_1 \text{BLAME} + \alpha_2 \text{AGE} + \sum_{l=1}^{8} \beta_l \Delta + \Gamma_i' \gamma + \epsilon_X, \quad (1)$$

where $P(X_i = 1)$ is the probability that choice X made by individual i is utilitarian, that is, saves the highest number of lives, and ϵ_X is the error term. We consider treatment-specific attributes (BLAME and AGE treatments), choice set-specific attributes, and individual characteristics Γ_i of respondent i. Concerning the choice set-specific attributes, we include the difference between the number of passengers and pedestrians, that is, Δ, with $\Delta = -4, -3, -2, -1, 1, 2, 3, 4$ (note that no utilitarian choices were possible for $\Delta = 0$ and were thus excluded). Negative (positive) differences imply that the pedestrians (passengers) are in the majority.

For the probability that individual i is utilitarian, that is, always makes the utilitarian choice in the choice sets in which this is possible, we estimate the following equation:

$$\text{logit}(P(Y_i = 1)) = \alpha_1 \text{BLAME} + \alpha_2 \text{AGE} + \Gamma_i' \gamma + \epsilon_i, \quad (2)$$

where we drop all choice set-specific attributes, as we consider all choices of an individual. The term $P(Y_i = 1)$ denotes the probability that individual i is utilitarian.

Table 4 presents the results. In column (1), we report the odds ratios, that is, the exponentiated regression coefficients, for Equation (1), and in column (2) we report the odds ratios for Equation (2). We cluster standard errors at the individual level in the regression in column (1). The reference group for BLAME and AGE is the CONTROL treatment group, and for the difference Δ, the reference group is $\Delta = 1$.

We find that individual characteristics do not have significant effects in either regression, except for the variable "Heard about the dilemma before", which is significant at the 10% level. Therefore, we conclude that individual characteristics do not play a huge role in determining what choices an individual makes. Our results are largely in line with those of Awad et al. (2018) who find that most individual variations have no sizable impact on moral preferences for dilemma situations involving autonomous vehicles. However, our results differ from theirs in that we do not find significant differences when it comes to gender.[12] In this regard, our study also contrasts with results obtained by Baez et al. (2017) who find that utilitarian responses to (traditional) moral dilemmas were less frequent in women, although the differences had very small effect sizes, potentially due to statistical power. Similarly, Fumagalli et al. (2010) and Friesdorf et al. (2015) found that men showed a stronger preference for utilitarian over deontological judgments than women.

In contrast to individual characteristics, (one of) the treatments and the attributes of the choice sets have much larger (and mostly strongly statistically significant) effects on the two outcome variables. As can be seen from columns (1) and (2) in Table 4, there is no significant difference between the BLAME and the CONTROL treatment, as also

TABLE 4 Logit regression of certain choices on treatments, choice set-specific attributes and individual characteristics.

	(1) Utilitarian choices	(2) Being utilitarian
BLAME	0.883	0.967
	(0.0792)	(0.155)
AGE	0.782***	0.443***
	(0.0627)	(0.0822)
$\Delta = -4$ (pedestrians in majority)	11.08***	
	(1.426)	
$\Delta = -3$	10.65***	
	(1.352)	
$\Delta = -2$	11.54***	
	(1.506)	
$\Delta = -1$	9.549***	
	(1.139)	
$\Delta = 2$ (passengers in majority)	1.228***	
	(0.0576)	
$\Delta = 3$	1.769***	
	(0.120)	
$\Delta = 4$	2.448***	
	(0.236)	
Male	1.035	0.944
	(0.0765)	(0.141)
Age	0.994	0.994
	(0.00346)	(0.00715)
University education	1.052	0.937
	(0.0908)	(0.165)
Personal monthly income (in TSEK)	1.001	1.007
	(0.00252)	(0.00470)
Children	1.004	1.069
	(0.0861)	(0.187)
Unemployed	1.292	1.714
	(0.354)	(0.821)
Retired	0.854	0.839
	(0.105)	(0.214)
Student	1.105	1.570
	(0.172)	(0.445)
Driver's license	1.042	0.817
	(0.132)	(0.196)
Car ownership	0.891	0.990
	(0.0908)	(0.210)
Commute to work by car	1.055	0.969
	(0.0936)	(0.178)

TABLE 4 (Continued)

	(1) Utilitarian choices	(2) Being utilitarian
Commute to work by foot	1.157	0.961
	(0.128)	(0.220)
Heard about dilemma	0.859**	0.984
	(0.0661)	(0.149)
Observations	10,460	1308
Pseudo R^2	0.178	0.028

Note: Odds ratios are reported. Standard errors are in parentheses and clustered by individual in column (1). The sample in column (1) contains all choice sets in which choices were non-missing and for which utilitarian choices were possible. Δ gives the difference between # of passengers and # of pedestrians.
*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

indicated by Table 3. However, the presence of children in the scenario (AGE treatment) has a significantly negative impact on the probability of making a utilitarian choice, and an even stronger one on the probability of being utilitarian (as compared to the CONTROL treatment). For example, in the AGE treatment, the odds ratio of sparing the higher number of lives decreases, ceteris paribus, by a factor of 0.782 as compared to the CONTROL treatment for utilitarian choices, and by a factor of 0.443 for being utilitarian. Strong predictors of the odds ratio of making a utilitarian choice are also a higher difference between the number of passengers and the number of pedestrians, especially the negative difference between the number of passengers and pedestrians. In particular, the larger the difference in the numbers of people in the two groups, the more likely is it that a respondent makes a utilitarian choice. However, there is a strong asymmetry between positive and negative differences (always as compared to the benchmark of the difference being one). If pedestrians are in the majority, this strongly increases the odds ratio of sparing the higher number of lives among the two groups. The regressions thus show that moral choices are highly contextual but that individual characteristics do not explain the difference in peoples' choices.

6 | RELATIVE VALUES OF LIVES

In this section, we proceed with the econometric analysis in order to show how respondents trade off the lives of passengers and pedestrians in the various treatments, both at the margin and at the point where they are indifferent between choosing "Continue forward" and "Swerve". To this end, we capture moral preferences by computing the average marginal rate of substitution between the lives of passengers and pedestrians. Doing so allows us to make statements such as the following: respondents in our sample are, on average and at the margin, willing to sacrifice x pedestrians in order to spare one (more) passenger (in a certain treatment). If the marginal rate of substitution is one, people do not, at the margin, have a preference for sparing a pedestrian over a passenger or the other way around. In other words, they value the lives of an additional pedestrian or an additional passenger equally. However, we also need to take into account that even if the lives of pedestrians and passengers may be valued equally at the margin, the respondent can have different preferences for the alternatives "Continue forward" and "Swerve" per se or be biased in her decision toward one of these two alternatives.

6.1 | Discrete choice model

As described earlier, all respondents were asked to choose their preferred options in $T = 10$ choice sets ($t = 1, 2, ..., 10$), where each choice set consisted of $J = 2$ alternatives. The alternatives j were either that the car is programmed to continue forward on the road ($j = C$), which implies certain death for one to five passengers, or swerve ($j = S$), whereby one to five pedestrians lose their lives. For simplicity, a linear random utility of respondent n from alternative j in choice set t is assumed and given by:

$$U_{njt} = \beta_{0nj} + \beta_{1n} \times \text{passengers}_{jt} + \beta_{2n} \times \text{pedestrians}_{jt} + \epsilon_{njt}, \tag{3}$$

where the β's are the coefficients to be estimated, and passengers_{jt}, $\text{pedestrians}_{jt} \in \{0, 1, 2, 3, 4, 5\}$ denote the numbers of passengers and pedestrians that are sacrificed in alternative j in choice set t. We set $\beta_{0nS} = 0$, and β_{0nC} is the alternative-specific constant (ASC) of the alternative "Continue forward". The coefficients β_{1n} and β_{2n} reflect the disutility that respondent n derives from the loss of (the number of) passengers and pedestrians, respectively. The systematic and observable part of the utility function $V_{njt} = \beta_{0nj} + \beta_{1n} \times \text{passengers}_{njt} + \beta_{2n} \times \text{pedestrians}_{njt}$ of the Random Utility Model (RUM) is followed by the random component ϵ_{njt}, which is assumed to be i.i.d. Gumbel distributed over people, choice sets t and alternatives j. Based on the RUM, the alternative $i = C, S$ is chosen over alternative $k \neq i$ if $U_{nit} > U_{nkt}$.

Assuming homogeneous preferences, respondent n's probability of choosing alternative i conditional on the vector β_n is given by the logit formula:

$$L_{ni} = \frac{e^{V_{nit}/\sigma}}{\sum_{j=C,S} e^{V_{njt}/\sigma}}. \tag{4}$$

Note that the coefficients cannot be separately identified from the scale parameter σ of the Gumbel distribution. Assuming that the coefficients are constant over choice situations and that respondents have stable preferences, we consider that each respondent makes several choices, in which case the probability of choosing alternative i conditional on β_n becomes a product of logit formulas:

$$L_{ni} = \prod_{t=1}^{T} \frac{e^{V_{nit}/\sigma}}{\sum_{j=C,S} e^{V_{njt}/\sigma}}. \tag{5}$$

Here, $i = i_1, i_2, ..., i_T$ represents a sequence of alternatives, one for each choice set. This is the standard logit model for panel data. However, it is known to be restrictive. One restriction is the limited ability to account for unobserved heterogeneity in taste. Therefore, we apply a Random Parameter Logit (RPL) model and estimate the simple logit model only for comparative purposes. In the RPL model, the coefficient vector β_n varies in the population with density $f(\beta)$. The RPL probability of choosing the sequence of alternatives i can then be expressed as:

$$P_{ni} = \int \prod_{t=1}^{T} \frac{e^{V_{nit}/\sigma}}{\sum_{j=C,S} e^{V_{njt}/\sigma}} f(\beta) d\beta. \tag{6}$$

The distribution of the coefficients can take various forms. To constrain the sign to be the same for every decision-maker in our application, the coefficients β_{1n} and β_{2n} are assumed to be triangularly distributed with the constraint that the mean equals the spread. Using the triangular distribution yields a negative coefficient, that is, there is a welfare loss from sacrificing passengers/pedestrians. No respondent thus derives positive utility from sacrificing anyone (which could be different for other distributions).[13] The ASC β_{0nC} is assumed to be normally distributed (see Train, 2009, for more details about the RPL model and simulated maximum likelihood methods that are used to estimate the model).

Once the model is estimated, the average marginal rate of substitution (MRS) between the loss of passengers and pedestrians can be computed by taking the total derivative of the observable part of the utility function and setting it equal to zero:

$$\text{MRS} = -\frac{d \text{ passengers}}{d \text{ pedestrians}}\bigg|_{dV=0} = \frac{\beta_2}{\beta_1}. \tag{7}$$

The MRS reflects the number of passengers that the decision-maker is willing to let go of in order to save one more pedestrian. Taking the ratio of the coefficients, that is, the ratio of the marginal utilities, the scale parameter drops out, and this in turn facilitates the comparison between the treatment groups.

6.2 | RPL results

Table 5 presents the results of the RPL model estimated for the three treatment groups. The models are estimated with Nlogit 6.0 using simulated maximum likelihood with Halton draws with 500 replications. There are no qualitative

TABLE 5 Logit model results.

	RPL model					
	CONTROL		BLAME		AGE	
	Coefficient	sd	Coefficient	sd	Coefficient	sd
Passengers	−1.86628***	1.86628***	−1.68293***	1.68293***	−1.31970***	1.31970***
	(0.14638)	(0.14638)	(0.12280)	(0.12280)	(0.10268)	(0.10268)
Pedestrians	−1.96528***	1.96528***	−1.44586***	1.44586***	−1.90861***	1.90861***
	(0.16912)	(0.16912)	(0.10815)	(0.10815)	(0.17724)	(0.17724)
ASC (β_{0nC})	6.20859***	8.44683***	2.53249***	9.02330***	5.93200***	4.83292***
	(0.63287)	(0.76472)	(0.40206)	(0.71441)	(0.58807)	(0.50367)
# respondents	445		441		431	
# responses	4342		4325		4148	
Log-likelihood	−1024		−1213		−769	
AIC	2057		2435		1547	

	Simple logit model (clustered standard errors)		
	CONTROL Coefficient	BLAME Coefficient	AGE Coefficient
Passengers	−0.40451***	−0.32403***	−0.48557***
	(0.03168)	(0.02522)	(0.04746)
Pedestrians	−0.38183***	−0.27299***	−0.51641***
	(0.03231)	(0.02224)	(0.05550)
ASC	1.36642***	0.54906***	2.18640***
	(0.10800)	(0.09323)	(0.15060)
# respondents	445	441	431
# responses	4342	4325	4148
Log-likelihood	−2079	−2702	−1279
AIC	4164	5411	2565

Note: Standard errors in parentheses.

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

differences between the treatments in terms of sign and significance of the coefficients. Furthermore, as expected, the first two coefficients (β_1 and β_2) are negative, meaning that people suffer a disutility from sacrificing members of either group, be they passengers or pedestrians.[14]

The ASC captures the average impact on utility from choosing one of the alternatives, of factors that are not included in the model as attributes. The positive ASC suggests that, on average, respondents prefer the self-driving car to be programmed to continue forward rather than swerve. This result could be due to several reasons. First, if respondents are rational, they might, on average, have an intrinsic preference for sparing the lives of pedestrians rather than those of passengers, for example, because they think that they are more likely to end up in such situations as a pedestrian than a passenger and would thus benefit from such a decision rule. Second, if individuals are not rational, their choices may be biased, for various reasons, toward the "Continue forward" choice. One such bias could be that respondents think that it is rather the passengers' fault than that of the pedestrians that the dilemma situation arises in the first place, since without the car on the road, no one would have to die. Another bias could be that respondents are reluctant toward the algorithm actively making a decision that causes loss of lives; they therefore prefer the car not to do anything, which means to stay put irrespective of the number of fatalities. Another closely related bias is the so-called status quo bias (Samuelson & Zeckhauser, 1988), where the current path acts as a reference point (or default state), and deviations from it are associated with a loss. In our setting, respondents could have perceived the choice "Continue

forward" as the default state. There is mixed evidence that such biases play a role in trolley dilemmas. For instance, as shown by Navarrete et al. (2011), people tend to be less utilitarian when a decision requires action, for example, pushing a man on the track whose body weight can stop the trolley and thus save the lives of the five on board of the trolley. By contrast, Frank et al. (2019) show that neither action nor status quo bias seem to play a role in trolley dilemmas with autonomous vehicles (in their paper, a change in the default path did not change respondents' decisions). In what follows, we will refer to the positive ASC for the choice "Continue forward" as "aversion to swerve" (or "swerve aversion").

Moreover, note that using the mean and standard deviation of the ASC, we can calculate the share of respondents that prefer to choose "Swerve" instead of "Continue forward", using the normal transformation.[15] We find that there is a non-negligible share of the respondents who prefer to swerve. In the CONTROL treatment, we have 23% of respondents that prefer this alternative, and for the BLAME and AGE treatment, the corresponding figures are 39% and 11%.

All coefficients in the RPL model are significant at the 1% level. Since the coefficients of the model are normalized by the variance of the unobserved factors (σ), there is no point in directly comparing the coefficients between the treatment groups. Rather, we calculate the MRS by taking the ratio of the coefficients. Which in turn facilitates the comparison between the treatment groups, as the scale parameter (σ) drops out. For comparison, and as a robustness check, we also present a logit model with fixed coefficients and clustered standard errors in the lower panel of Table 5. The first thing to note is that the Akaike information criterion is much higher for the logit model than for the RPL model, which supports the use of the RPL model. Qualitatively, however, the models produce similar results in terms of sign and significance.

Table 6 presents the MRS of the average decision-maker based on the RPL model. We find that there are significant differences between the three treatments at the 5% or 1% significance level.[16]

In the CONTROL treatment, people are prepared to roughly let go of one passenger (1.053 to be precise) to save the life of one additional pedestrian (while keeping their utility constant). When pedestrians are to blame for the accident, about 0.9 passengers are sacrificed to save the life of one additional pedestrian. This implies that a pedestrian's life is now valued less than that of a passenger. When children are involved as pedestrians, people are willing to sacrifice about 1.4 passengers to save one child on the road; they thus value the child's life (much) more than that of the adult.[17]

In summary, our results show that, on average, respondents in the CONTROL treatment derive the same marginal utility from saving either one additional pedestrian or one additional passenger. Furthermore, we find that who is to blame for the accident is important to people when it comes to the ethical decision regarding whom to save. Similarly, the age of the persons that can be saved is an important criterion in this decision. Next, we will focus on the aversion to swerve, which is captured by the ASC. This allows us to look at the trade-off beyond the marginal valuation of saving either one additional pedestrian or one additional passenger.

6.3 | The total cost of saving pedestrians

So far, we have looked at the trade-off between saving pedestrians' or passengers' lives *at the margin*, that is, when we are thinking of sacrificing (or saving) one *additional* pedestrian, while holding utility constant. We saw that respondents

TABLE 6 Marginal rates of substitution across treatments, based on the random parameter logit model.

	CONTROL	BLAME	AGE
MRS	1.053 (0.061)	0.859 (0.052)	1.446 (0.110)
95% confidence interval	[0.933; 1.172]	[0.758; 0.960]	[1.230; 1.663]
CONTROL versus BLAME[a]	$Z = 2.42$**		
CONTROL versus AGE	$Z = -3.12$***		
BLAME versus AGE	$Z = -4.824$***		

Note: Standard errors in parentheses, estimated using the delta method.

[a]Test statistic for equality of marginal rates of substitution across treatment groups.

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

needed to be compensated in this case, on average, with one passenger saved in the CONTROL treatment, less than one passenger saved in the BLAME treatment, and more than one passenger saved in the AGE treatment.

In the next step, we calculate (in terms of the number of passengers that need to be saved) the total utility cost of sacrificing one or more pedestrians (or the other way round), taking into account the aversion to swerve, as captured by the ASC. For this purpose, we calculate when the average respondent is indifferent between the choices "Continue forward" and "Swerve".

To this end, consider the difference in the observable part of the utility between the choices "Continue forward" and "Swerve" for the average respondent in our sample, which we define as ΔV:

$$\begin{aligned}\Delta V &= V_C - V_S \\ &= \beta_{0C} + \beta_1 \text{passengers}_C + \beta_2 \text{pedestrians}_C - \beta_1 \text{passengers}_S - \beta_2 \text{pedestrians}_S \\ &= \beta_{0C} + \beta_1 \text{passengers}_C + \beta_2 0 - \beta_1 0 - \beta_2 \text{pedestrians}_S \\ &= \beta_{0C} + \beta_1 \text{passengers}_C - \beta_2 \text{pedestrians}_S,\end{aligned} \tag{8}$$

where, for example, passengers_C is strictly positive for the alternative "Continue forward" (because the passengers' lives are sacrificed, i.e., there is a disutility from sacrificing passengers), in which case passengers_S is zero for the other alternative (because the pedestrians' lives are spared and thus there is no disutility from sacrificing them).

Setting the above equation equal to zero (implying that the average respondent is indifferent between the two alternatives) and solving it for the number of passengers yields:

$$\text{passengers}_C = \frac{-\beta_{0C} + \beta_2 \text{pedestrians}_S}{\beta_1} = \frac{-\beta_{0C}}{\beta_1} + \text{MRS} \times \text{pedestrians}_S. \tag{9}$$

Plugging in the results of the RPL model in Table 5 into Equation (9) and setting $\text{pedestrians}_S = 1$ yields the results found in Table 7. These results thus indicate how many passengers need to be present in a scenario for the average respondent to be indifferent between the choices "Continue forward" and "Swerve" when one pedestrian is present.

Notably, taking the "swerve aversion" into account increases the number of passengers that the decision-maker is willing to sacrifice in order to save pedestrians. For example, the average respondent is willing to sacrifice about four passengers to save the first pedestrian on the road in the CONTROL treatment. This gives a more realistic but complex understanding of the respondents' choice. On the one hand, the marginal valuation of saving a passenger's life is the same as saving a pedestrian's life in the CONTROL treatment. On the other hand, respondents have a relatively strong swerve aversion, which makes the average respondent behave less utilitarian. In the CONTROL treatment, for example, the respondents will choose "Continue forward" and thus spare the only pedestrian on the road as long as there are four or less than four passengers in the car. With five or more passengers in the car, the decision would be made in favor of the passengers, while another pedestrian on the road would again revert the decision back to "Continue forward".

Respondents are, on average, substantially less willing to sacrifice passengers if the pedestrians are to blame for the dilemma (about two passengers for the first pedestrian), and substantially more passengers if the pedestrians are children (about six passengers for the first pedestrian).[18]

To conclude, once we estimate the respondents' preferences, we observe that respondents are not utilitarian on average. The swerve aversion makes the average respondent behave less utilitarian, although she values lives at the

TABLE 7 Number of passengers that need to be sacrificed in order to save one pedestrian, based on the random parameter logit model.

	CONTROL	BLAME	AGE
passengers_C	4.380 (0.061)	2.364 (0.186)	5.941 (0.346)
95% confidence interval	[3.894; 4.865]	[2.000; 2.728]	[5.263; 6.619]
CONTROL versus BLAME[a]	$Z = 10.299$***		
CONTROL versus AGE	$Z = -4.443$***		
BLAME versus AGE	$Z = -9.106$***		

Note: Standard errors in parentheses.

[a]Test statistic for equality of passengers_C across treatment groups.

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

margin equally (at least in the CONTROL treatment). Our results thus indicate that there is a certain individual threshold for the difference between the number of passengers and pedestrians beyond which people are more utilitarian.

7 | DISCUSSION

While our study dives into more detail than the previous literature in several respects, there are scenarios and factors that we could not address. First, the groups of passengers and pedestrians in our study are homogeneous with respect to age and other characteristics. All passengers and pedestrians were depicted in a gender-neutral way, and we have not asked respondents to make decisions between groups that consisted of people of different genders or ages. While one would hope that gender does not make a difference in people's moral decisions, it cannot be taken as a given. For example, Skulmowski et al. (2014) found a general tendency toward sacrificing male individuals in a virtual reality setting inducing the trolley dilemma.

Second, we examined deterministic scenarios for which it was clear that either the group of passengers or the group of pedestrians would face certain death. In reality, passengers and pedestrians might not necessarily die in a dilemma situation but be vulnerable and injured to a different degree, depending, for example, on the size of the car and its safety features. An algorithm for self-driving cars could potentially integrate such considerations, trying to minimize the probability that someone will be harmed, or minimize overall harm (e.g., two severely wounded people might be perceived as a better outcome than one person killed) rather than maximize the number of lives saved. In this sense, our study delineates the most extreme situation that can arise. Meder et al. (2018) investigate dilemma situations with self-driving vehicles in the presence of risk (where probabilities of dying in case of a certain action exist and are known) or uncertainty (where these probabilities exist but are unknown). They find that subjects' decision preferences and moral judgments varied considerably under risk, but less so under uncertainty. Furthermore, in their study, staying in the lane and performing an emergency stop was generally considered a reasonable default, even in cases when this action did not minimize expected losses.

8 | CONCLUSION

In some countries, governments have installed ethics committees for making recommendations for the programming of future autonomous vehicles, particularly in dilemma situations. For example, the German government mandated an ethics committee consisting of 14 scientists and experts from the fields of ethics, law, and technology. This committee stated and in fact started its work with the instruction that any discrimination against people in such dilemma situations based on their sex and age, or any other characteristics, shall not be permissible. The experts also spoke out against using utilitarian ethics in dilemma situations. At the same time, they acknowledge that it might be ethically defendable to minimize the number of injured or killed people if this lowers the risk for everyone ex-ante to the same extent, that is, when it is not clear beforehand that a particular person will become a victim. Such situations are, however, hard to conceive. One example could be that it is permissible to discriminate against passengers in cars, as they are much less vulnerable to severe injuries than pedestrians who are basically unprotected in a crash.

This paper has analyzed the preferences of a representative sample of the Swedish population regarding the programming of algorithms for self-driving cars. Our main conclusion is that respondents' preferences are largely at odds with the recommendations of the above-mentioned ethics committee. In particular, our results suggest that people do think that factors like the age of the people involved, the degree to which they can be blamed for causing a dilemma situation, and the ratio of the number of people in the two groups (passengers vs. pedestrians) should be considered when programming self-driving cars. We find that people do, on average and at the margin, value the lives of passengers and pedestrians equally when people in the two groups are of equal age and no social norm is violated, but this does not hold when blame or children are involved. In addition, there is a strong aversion to sacrificing pedestrians. Most of these results thus speak against applying utilitarian ethics in dilemma situations, unless it is in situations in which one group significantly outweighs the other one in terms of its sheer number. Experts' and laymen's judgments regarding the programming of algorithms in self-driving cars seem to lie too far apart at the moment to reach a societal consensus. Should the actual programming of autonomous vehicles go against people's preferences, one may thus expect that the widespread adoption of these cars will be hampered.

Why people have such a strong aversion to sacrificing pedestrians remains an interesting question for future research. One reason for this could be that more responsibility for the occurrence of the dilemma situation is implicitly attributed to the passengers, as there would have been no dilemma at all if the car had not been on the road in the first place. Another possible explanation is that people find it wrong to change the car's direction, as they might think that they get more actively involved in the moral trade-off between passenger and pedestrian deaths when doing so. More research is needed to understand the motivation behind people's trade-offs in such dilemma situations.

ACKNOWLEDGMENTS

We thank Maria Andreasson and Johan Martinsson of the Laboratory of Opinion Research (LORE) at the University of Gothenburg for their help in designing and carrying out the survey. We also thank Stefan Bauernschuster, Michael Grimm, Vera Huwe, Francois Laisney, Johann Graf Lambsdorff, and seminar participants at the University of Passau for valuable comments and suggestions, and we thank Nicholas Eveneshen for proof-reading. The survey was financed by the University of Augsburg and the University of Gothenburg.

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

None.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon request.

ORCID

Wolfgang Habla https://orcid.org/0000-0003-1164-0962
Peter Martinsson https://orcid.org/0000-0002-1146-9248

ENDNOTES

[1] Algorithms that have been confronted and trained with dilemma situations have a thousand-fold higher computing capacity than the human brain. Therefore, the overall number of accidents and fatalities is generally expected to decrease with the widespread adoption of self-driving cars. However, it is unlikely that accidents and the occurrence of dilemma situations can be completely eliminated, as unexpected situations can happen anytime.

[2] In addition, self-driving cars may also lead to more participation in social life for less mobile individuals who are not able to drive themselves (Curl & Fitt, 2019).

[3] It is often claimed that trolley cases are of little or no relevance to the ethics of self-driving vehicles. Keeling (2019) identifies four arguments for this view and shows how one can reject all of those. He argues that trolley-style problems still inform the ethics of self-driving vehicles even if we will never see them in the real world. One of the reasons why they are relevant is "the not unreasonable hope [...] that the relation between the properties of acts and their moral permissibility in *ideal* cases is relevant to the moral permissibility of acts which instantiate those same properties in more noisy real-world cases" (Keeling, 2019, p. 4).

[4] One could also define a choice to be utilitarian if it saves the highest number of life-years. This definition would, however, not be applicable in our study, as we have not specified people's age or life expectancy.

[5] As mentioned in footnote 4, utilitarianism could be defined in a way that it spares the group with the highest number of life-years saved. If this applies, then saving children could be in line with utilitarian choices in many of the dilemma situations considered in this study, since children have more years to live than the adult passengers.

[6] In Study 2 of Bonnefon et al. (2016), there are scenarios with one passenger versus one, two, five, 20 or 100 pedestrians.

[7] To keep the survey as short as possible, we did not explain how the cars can tell adults and children apart. This differentiation can be inferred from the height of pedestrians or the weight passengers exert on the seats.

[8] A concrete barrier on both sides of the road as displayed in the pictorial representations of the different treatments precludes any other options, like swerving to the right. We made clear in the survey that this course of action would result in the same outcome as continuing forward.

[9] The reason why we only specified the children's age is that children below the age of 10 are supposed to be accompanied by their parents or other adults in Sweden. If we had not specified the age, this could have confused respondents as to why the children are unaccompanied.

[10] They could have made a maximum of 13,170 choices but left 355 choice sets unanswered.

[11] Similarly, Figure A1 in the Appendix illustrates the shares of "Continue forward" choices depending on this difference.

[12] Awad et al. (2018) also found significant differences when it comes to religiosity. We did not ask about religiosity, because Sweden is known to be a very secular country.

[13] We also used the log-normal distribution, but the results did not converge for this specification for two of the three treatments. Using the normal distribution yielded relatively large standard errors and results that are quite off compared to what one would expect. Furthermore, the RPL model with the triangular distribution is relatively close to the standard logit model (where we used clustered standard errors to incorporate repeated choices).

[14] Interaction terms were included in an extended model to test if people that usually commute by car have a higher disutility when passengers are sacrificed, and if people that usually commute as pedestrians have a higher disutility when pedestrians are sacrificed. The interaction terms were—with only one exemption—statistically insignificant and were therefore dropped in the final analysis.

[15] For example, in the CONTROL treatment $z = \frac{x-\mu}{\sigma} = \frac{0-6.21}{8.45} = -0.73$. The probability to swerve is thus $P(Z < -0.73) = 0.23$.

[16] The parametric test is based on independent samples assuming a normal distribution where $Z = (MRS_i - MRS_j)/\sqrt{\mathrm{SE}^2_{MRS_i} - \mathrm{SE}^2_{MRS_j}}$ and where indices i and j refer to different treatments. We also used the non-parametric test in Poe et al. (2005) to confirm our results. All compared populations are found to be statistically different from each other at the 1% significance level.

[17] To test for the robustness of our results, we also estimated models that capture heterogeneity in mean to predict MRS based on Swedish population average values for the variables male, age, personal monthly income, and university education. Unfortunately, when we add heterogeneity in mean, the preservation of the sign is not imposed anymore, which is why we do not take an interest in the actual MRS values from these models but only focus on the issue of misrepresentation. In any case, these models confirm that the small misrepresentation does not affect the estimated MRS in any considerable way.

[18] The statistical test is based on independent samples assuming a normal distribution. We also used the non-parametric test in Poe et al. (2005) to confirm our results. All compared populations are found to be statistically different from each other at the 1% significance level.

REFERENCES

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

Baez, S., Flichtentrei, D., Prats, M., Mastandueno, R., García, A. M., Cetkovich, M., & Ibáñez, A. (2017). Men, women...who cares? A population-based study on sex differences and gender roles in empathy and moral cognition. *PLoS One*, *12*(6), e0179336. https://doi.org/10.1371/journal.pone.0179336

Baltussen, R., Stolk, E., Chisholm, D., & Aikins, M. (2006). Towards a multi-criteria approach for priority setting: An application to Ghana. *Health Economics*, *15*(7), 689–696. https://doi.org/10.1002/hec.1092

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Cappelen, A., & Norheim, O. (2005). Responsibility in health care: A liberal egalitarian approach. *Journal of Medical Ethics*, *31*(8), 476–480. https://doi.org/10.1136/jme.2004.010421

Carlsson, F., Daruvala, D., & Jaldell, H. (2010). Preferences for lives, injuries, and age: A stated preference survey. *Accident Analysis & Prevention*, *42*(6), 1814–1821. https://doi.org/10.1016/j.aap.2010.05.002

Covey, J., Robinson, A., Jones-Lee, M., & Loomes, G. (2010). Responsibility, scale and the valuation of rail safety. *Journal of Risk and Uncertainty*, *40*(1), 85–108. https://doi.org/10.1007/s11166-009-9082-0

Cropper, M., Ayede, S., & Portney, P. (1994). Preferences for life saving programs: How the public discounts time and age. *Journal of Risk and Uncertainty*, *8*(3), 243–265. https://doi.org/10.1007/bf01064044

Curl, A., & Fitt, H. (2019). Will driverless cars be good for us? Now is the time for public health to act together with urban and transport planning. *Journal of Global Health*, *9*(2), 020303. https://doi.org/10.7189/jogh.09.020303

Dolan, P., & Tsuchiya, A. (2005). Health priorities and public preferences: The relative importance of past health experience and future health prospects. *Journal of Health Economics*, *24*(4), 703–714. https://doi.org/10.1016/j.jhealeco.2004.11.007

Dolan, P., & Tsuchiya, A. (2009). The social welfare function and individual responsibility: Some theoretical issues and empirical evidence. *Journal of Health Economics*, *28*(1), 210–220. https://doi.org/10.1016/j.jhealeco.2008.10.003

Edlin, R., Tsuchiya, A., & Dolan, P. (2012). Public preferences for responsibility versus public preferences for reducing inequalities. *Health Economics*, *21*(12), 1416–1426. https://doi.org/10.1002/hec.1799

Einhorn, J., Andersson, I., Carlson, L., Hallerby, N., Krook, C., Lindqvist, B., & Östh, R. (1995). "Vårdens såvra val". Retrieved from https://www.regeringen.se/contentassets/6c4cb9f4c3ef4296b68ea7c6cefbf1d2/del-1-kap.-1-t.o.m.-kap.-8-vardens-svara-val. Online; accessed: 2023-08-22.

Engisch, K. (1930). *"Untersuchungen über Vorsatz und Fahrlässigkeit im Strafrecht"*. O. Liebermann.

Fleetwood, J. (2017). Public health, ethics, and autonomous vehicles. *American Journal of Public Health*, *107*(4), 532–537. https://doi.org/10.2105/ajph.2016.303628

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

Frank, D.-A., Chrysochou, P., Mitkidis, P., & Ariely, D. (2019). Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports*, *9*(1), 13080. https://doi.org/10.1038/s41598-019-49411-7

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, *41*(5), 696–713. https://doi.org/10.1177/0146167215575731

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., Lucchiari, C., Consonni, D., Nordio, F., Pravettoni, G., Cappa, S., & Priori, A. (2010). Gender-related differences in moral judgments. *Cognitive Processing*, *11*(3), 219–226. https://doi.org/10.1007/s10339-009-0335-2

Gu, Y., Lancsar, E., Ghijbem, P., Butler, J., & Donaldson, C. (2015). Attributes and weights in health care priority setting: A systematic review of what counts and to what extent. *Social Science & Medicine*, *146*, 41–52. https://doi.org/10.1016/j.socscimed.2015.10.005

Internet Encyclopedia of Philosophy. (2019). "Consequentialism". Retrieved from https://www.iep.utm.edu/conseque/. Online; accessed: 2019-09-30.

Johansson-Stenman, O., Mahmud, M., & Martinsson, P. (2011). Saving lives vs. life-years in rural Bangladesh: An ethical preferences approach. *Health Economics*, *20*(6), 723–736. https://doi.org/10.1002/hec.1627

Johansson-Stenman, O., & Martinsson, P. (2008). Are some lives more valuable? An ethical preferences approach. *Journal of Health Economics*, *27*(3), 739–752. https://doi.org/10.1016/j.jhealeco.2007.10.001

Keeling, G. (2019). *"Why trolley problems matter for the ethics of automated vehicles"*. Science and Engineering Ethics.

Martinsson, J., Andreasson, M., & Markstedt, E. (2017). *Technical report citizen panel 24 – [2017]*. University of Gothenburg, LORE.

Meder, B., Fleischhut, N., Krumnau, N.-C., & Waldmann, M. R. (2018). How should autonomous cars drive? A preference for defaults in moral judgments under risk and uncertainty. *Risk Analysis*, *34*, 311–321.

Mill, J. S. (1998). *Utilitarianism [1861]*. Oxford University Press.

Navarrete, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2011). Virtual morality: Emotion and action in a simulated three-dimensional trolley problem. *Emotion*, *12*(2), 364–370. https://doi.org/10.1037/a0025561

Olsen, J. A., Richardson, J., Dolan, P., & Menzel, P. (2003). The moral relevance of personal characteristics in setting health care priorities. *Social Science & Medicine*, *57*(7), 1163–1172. https://doi.org/10.1016/s0277-9536(02)00492-6

Poe, G. L., Giraud, K. L., & Loomis, J. B. (2005). Computational methods for measuring the difference of empirical distributions. *American Journal of Agricultural Economics*, *87*(2), 353–365. https://doi.org/10.1111/j.1467-8276.2005.00727.x

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*(1), 7–59. https://doi.org/10.1007/bf00055564

Skulmowski, A., Bunge, A., Kaspar, K., & Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: A virtual reality and eye tracking study. *Frontiers in Behavioral Neuroscience*, *8*, 426. https://doi.org/10.3389/fnbeh.2014.00426

Swann, W. B., Gomez, A., Dovidio, J. F., Hart, S., & Jetten, J. (2010). Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science*, *21*(8), 1176–1183. https://doi.org/10.1177/0956797610376656

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395–1415. https://doi.org/10.2307/796133

Train, K. (2009). *"Discrete choice methods with simulation"*. Cambridge University Press.

WHO (2018). "Global status report on road safety 2018".

Williams, A., & Cookson, R. (2000). Equity in health (chapter 35). In A. Culyer & J. Newhouse (Eds.), *Handbook of health economics volume 1 Part B*. Elsevier.

Zamir, E., & Medina, B. (2010). Threshold deontology and its critique. In *Law, economics, and morality*. Oxford University Press.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Habla, W., Kataria, M., Martinsson, P., & Roeder, K. (2024). Should it stay, or swerve? Trading off lives in dilemma situations involving autonomous cars. *Health Economics*, 1–23. https://doi.org/10.1002/hec.4802

APPENDIX A

TABLE A1 *p*-values of statistical tests for differences across treatment groups (pair-wise comparison across treatments).

	CONTROL vs. BLAME	CONTROL vs. AGE	BLAME vs. AGE
Male[b]	0.6303	0.3560	0.6572
Age[a]	0.6284	0.7917	0.4574
University education[b]	0.8657	0.3282	0.2530
Personal monthly income[a]	0.1995	0.5567	0.4879
Children[b]	0.7351	0.5856	0.8348
Unemployed[b]	0.4512	0.7468	0.6692
Retired[b]	0.3547	0.4547	0.8630
Student[b]	0.3473	0.2004	0.7292
Additional covariates:			
Driver's license[b]	0.7266	0.9311	0.6649
Car ownership[b]	0.5986	0.7547	0.8332
Commute to work by car[b]	0.6866	0.3749	0.6275
Commute to work by foot[b]	0.3929	0.4305	0.9511
Heard about dilemma before[b]	0.9735	0.2544	0.2419

[a]*t*-test for equal sample means.

[b]Equal proportions test.

	CONTROL	BLAME	AGE
$\Delta = -4$	0.872	0.739	0.932
$\Delta = -3$	0.870	0.738	0.918
$\Delta = -2$	0.890	0.736	0.935
$\Delta = -1$	0.865	0.704	0.913
$\Delta = 1$	0.350	0.516	0.153
$\Delta = 2$	0.396	0.556	0.209
$\Delta = 3$	0.456	0.616	0.338
$\Delta = 4$	0.517	0.704	0.432
$\lvert\Delta\rvert = 1$	0.627	0.620	0.568
$\lvert\Delta\rvert = 2$	0.678	0.663	0.627
$\lvert\Delta\rvert = 3$	0.702	0.689	0.684
$\lvert\Delta\rvert = 4$	0.758	0.727	0.767
Observations (# of choices)	3545	3543	3444

TABLE A2 Share of utilitarian choices depending on the treatment and the difference in the total number of people saved (Δ = number of passengers in the choice set minus number of pedestrians in the same choice set, i.e., $\Delta > 0$ implies that passengers are in the majority, while $\Delta < 0$ implies pedestrians are in the majority; $|\Delta|$ is the absolute value of the difference).

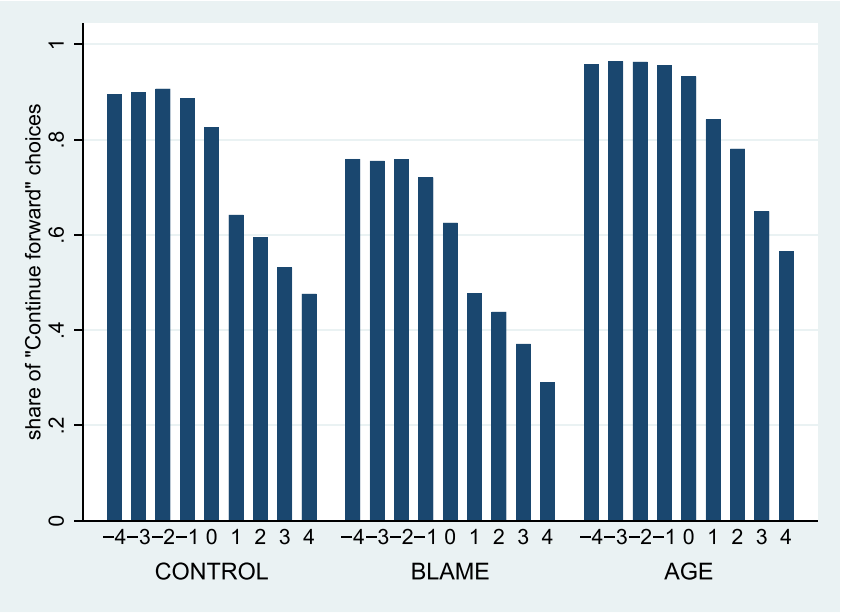


FIGURE A1 Shares of "continue forward" choices, depending on the treatment and the difference in the number of people that can be saved (= number of passengers in the choice set minus number of pedestrians in the same choice set, i.e., negative numbers indicate that pedestrians are in the majority, while positive numbers indicate that passengers are in the majority).