Issue 20(3), 2020, pp. 71-86
https://doi.org/10.18757/ejtir.2020.20.3.4515

TU Delft OPEN

EJTIR

ISSN: 1567-7141
http://ejtir.tudelft.nl/

Death by automation—Differences in weighting of fatalities caused by automated and conventional vehicles

Bing Huang[1]
Department of Engineering Systems and Services, Delft University of Technology, Netherlands.

Sander van Cranenburgh[2]
Department of Engineering Systems and Services, Delft University of Technology, Netherlands.

Caspar G. Chorus[3]
Department of Engineering Systems and Services, Delft University of Technology, Netherlands

Although Automated vehicles (AVs) are expected to have a major and positive effect on road safety, recent accidents caused by AVs tend to generate a powerful negative impact on the public opinion regarding safety aspects of AVs. Triggered by such incidents, many experts and policy makers now believe that paradoxically, safety perceptions may well prohibit or delay the rollout of AVs in society, in the sense that AVs will need to become much safer than conventional vehicles (CVs), before being accepted by the public. In this study, we provide empirical insights to investigate and explain this safety paradox. Using stated choice experiments, we show that there is indeed a difference between the weight that individuals implicitly attach to an AV-fatality and to a CV-fatality. However, the degree of overweighting of AV-fatalities, compared to CV-fatalities, is considerably smaller than what has been suggested in public opinions and policy reports. We also find that the difference in weighting between AV-fatalities and CV-fatalities is (partly) related to a reference level effect: simply because the current number of fatalities caused by AVs is extremely low, each additional fatality carries extra weight. Our findings suggest that indeed, AVs have to become safer—but not orders of magnitude safer—than CVs, before the general public will develop a positive perception of AVs in terms of road safety. Ironically, our findings also suggest that the inevitable occurrence of more AV-related road accidents will in time lead to a diminishing degree of overweighting of safety issues surrounding AVs.

Keywords: *automated vehicles; weighting of fatalities; reference level effects; stated choice experiments; road safety.*

[1] A: Jaffalaan 5, 2628BX Delft, The Netherlands T: +31 152 787 673 F: +31 152 787 673 E: b.huang-1@tudelft.nl
[2] A: Jaffalaan 5, 2628BX Delft, The Netherlands T: +31 152 786 957 F: +31 152 786 957 E: s.vancranenburgh@tudelft.nl
[3] A: Jaffalaan 5, 2628BX Delft, The Netherlands T: +31 152 788 546 F: +31 152 788 546 E: c.g.chorus@tudelft.nl

1. Introduction

The advent of Automated—or autonomous, or self-driving—vehicles (AVs) is generally believed to have positive effects on a variety of dimensions, such as road capacity (Shladover et al., 2012), emissions (Greenblatt & Saxena, 2015), travel time (de Almeida Correia et al., 2019), and road safety (Simonite, 2013; Sparrow & Howard, 2017). As for the latter aspect, AVs are expected to have a significant impact on decreasing the number of traffic accidents, as many car crashes are the result of human errors (Singh, 2015), which can be vastly reduced by the assistance of automation technology (Anderson et al., 2016; Fagnant & Kockelman, 2015). However, a considerable degree of societal anxiety exists regarding safety aspects of AVs; colloquially discussed in terms of anxiety and fear of "being killed by a robot". A report from the American automobile association (AAA, 2017) revealed that 78% of American drivers were afraid to ride in AVs. The dominating societal concern about this new technology is software hacking or misuses, according to a large-scale survey which contains 5,000 respondents from 109 countries (Kyriakidis et al., 2015). Other risks that the public is concerned about include hardware or software failure (Piao et al., 2016). Moreover, the public debate has given ample attention to potentially disturbing moral aspects of AVs, such as the necessity for AVs to make life-and-death decisions (Awad et al., 2018; Bonnefon et al., 2016; Shariff et al., 2017); this may lead to severe ethical conce rns and hesitates among planners and regulators. All these perceived risks regarding safety aspects can be big barriers standing in the way of AV mass adoption.

As a result of these social concerns, AVs seem to be subject to what may be called a *safety paradox*: although AVs are expected to substantially contribute to road safety (i.e., eliminating a large share of traffic accidents), concerns among the public (and as a result, policy makers) regarding AV safety issues may well hamper or delay the rollout of AVs. This paradox has found its way to public debates, as illustrated by the following two examples. Gill Prate, CEO of the Toyota Research Institute, stated that "even cutting the number of annual fatalities in half—saving 18,000 lives in the United States for example—would not be enough for AVs to win the public' trust" (Mervis, 2017). Amnon Shashua, a maker of AV-technology, even claims that a thousand-fold improvement in safety, compared to conventional vehicles (CVs), is needed for AVs to be accepted by the general public (Economist, 2018). Such discussions can be also found in academic papers. Nees (2019) claimed that although the safety level required for AVs is still unclear, the benchmark of being safer than average human drivers does not seem to be adequate. A very recent study conducted by Liu et al. (2019) attempts to answer this provocative question "how safe is safe enough for AVs?" They examined the relationship between risk frequencies (e.g., one fatality per one million population) and risk-acceptance rates (i.e., the percentage of people accepting presented risk scenarios). The results showed that AVs should be four to five times safer than human drivers in order to be tolerated by the public.

Given the important policy implications of this safety paradox—a delay in the introduction of AVs may in fact inadvertently lead to a failure to avoid thousands of traffic fatalities (Kalra & Groves, 2017), this paper attempts to put this apparent safety paradox to the empirical test, and also to find potential explanations for it. More specifically, we answer the research questions: *whether—and if so, why—the fatalities caused by AVs carry more weight to the general public, compared to the fatalities caused by CVs*. The starting point for our analysis is a recent stated choice (SC) study (Overakker, 2017) which positively answers the "whether" part of the above research questions; the study finds that AV-fatalities are weighted much more than CV-fatalities. Specifically, AV fatalities caused by software bugs are weighted around four times higher than fatalities caused by CVs; AV fatalities caused by software hacking are weighted even higher—5.5 times. These empirical findings are in line with what was reported in Liu et al. (2019), despite that these two studies used different approaches. In this research, we conduct two SC experiments. Our first experiment, called

experiment A, aims to replicate[4] the study by Overakker, to examine whether—and the extent to which—AV-fatalities are weighted more than CV-fatalities. The second experiment, called experiment B, is designed specifically to find explanations for the overweighting of AV-fatalities, compared to CV-fatalities[5]. More specifically, experiment B is designed in such a way that it allows us to answer the following question: *is the difference in the weighting between AV-fatalities and CV-fatalities caused by the intrinsic, qualitative differences between an AV and a human-operated CV? Or is it caused by the fact that, quantitatively speaking, the level of AV-related fatalities is currently so low that each additional fatality receives considerable extra weight?* Note that disentangling these two potential explanations is crucial, not just from a scientific point of view, but also from a policy point of view. For example, the first explanation would suggest that problematic safety perceptions surrounding AVs may persist at least during the near future; up until the point where humans have truly got used to interacting with AVs. While the second explanation would suggest that, ironically, the inevitable occurrence of more AV-related road accidents will in time lead to a diminishing degree of overweighting of safety issues surrounding AVs.

This paper aims to contribute to the literature on social acceptance of AVs by focusing on a very specific, but highly salient aspect of safety. We use SC experiments to derive and explain the differences in weight attached by citizens to AV- and CV-fatalities, rather than asking them directly about their safety perceptions concerning AVs (relative to CVs); to the best of our knowledge, this approach has not yet been used in the scholarly literature. The remainder of this paper is organized as follows: the next section introduces the experiments and data collection effort. Section 3 presents models and estimation results. Section 4 discusses the obtained findings and related policy implications, as well as limitations.

2. Experimental designs and data collections

2.1 Experimental designs

Experiments were designed to observe choices that respondents made between hypothetical scenarios in the form of policy packages for the AV era. Specifically, respondents were informed that the government was considering to develop a long-term transport policy to anticipate and facilitate the large-scale introduction of AVs. The background was set at year 2045, and participants were informed that about 50% of traffic would consist of full AVs, according to recent studies regarding AV predictions (Litman, 2019; Nieuwenhuijsen et al., 2018). The hypothetical scenarios were described in terms of the consequences of the policy packages, which were presented in choice tasks of three alternatives. Each alternative was described by four following attributes (attribute levels are in brackets):

- The number of fatalities per year caused by CVs (250, 300, 350, 400);
- The number of fatalities per year caused by technical failures (e.g., a software bug) of AVs (50, 100, 150, 200);
- The number of fatalities per year caused by a malicious act (e.g., software hacking) regarding AVs (0, 30, 60, 90);
- The average reduction in car travel time (-30%, -20%, -10%, 0%).

See Figure 1 for a detailed wording of these four attributes. The attribute levels of these fatalities were designed according to the current situation in the Netherlands—there are around 600

[4] Note that we do not aim to replicate the exact same experiment as Overakker's. His experiment focused on the social acceptance of AVs which included many AV-related attributes. This research aims to examine the difference in weighting between AV- and CV-fatalities, and to find explanations for the difference; details regarding experimental set-ups are provided in Section 2.

[5] Note that the experimental set-up in Overakker (2017) did not allow for studying the possible causes for the overweighting of AV-fatalities.

fatalities each year (SWOV, 2019), and all these fatalities are CV-related. It is uncertain whether the partial introduction of AVs would increase the fatal rate of CV-fatalities or not in the future, thus the attribute level of CV-fatalities was designed to range from 250 to 400, which roughly pivots around 300 (which is half of the current number of CV-related fatalities). Note that while the focus of the experiments is on AV- and CV-fatalities, car travel time was also included in order to help reduce the odds that respondents might add up all fatality numbers and then choose the smallest total. As for other possibly relevant criteria (e.g., cost, feasibility), respondents were informed that they were the same in every policy package.

A D-efficient design was applied to ensure a statistically efficient data collection (Rose & Bliemer, 2009). The priors were obtained by conducting a small pilot study (N=31). Eventually, twelve choice tasks were generated, and they were grouped into two blocks containing six tasks each. Each respondent was asked to complete one block.

2.2 Experiment treatment

As mentioned in the Introduction, there are two experiments designed in this study. These two experiments contained the same choice tasks. The only difference was the reference level provided to respondents. More specifically, before performing the choice tasks, respondents were requested to read an introduction page which contained reference levels of CV- and AV-fatalities. The details regarding the reference levels are described as follows.

- Experiment A

Experiment A provided real current levels in the context of the Netherlands: 600 fatalities per year caused by CVs, zero fatalities caused by AVs (either by technical failures or deliberate misuse), and a zero percent reduction in travel time. See Figure 1[6] for an example of how these current level-based reference points are visualized.

Q1. Suppose that these are the outcomes, in 2045, of 3 different policy packages. Which policy package would you vote for?

	Policy package 1	Policy package 2	Policy package 3	current situation
Average reduction in car travel time	-10%	-30%	0%	0%
*Fatalities caused by **conventional cars** (e.g. the driver not paying attention)*	350	250	400	600
*Fatalities caused by **technical failure of the AV** (e.g. a software bug)*	200	100	100	0
*Fatalities caused by **deliberate misuse of the AV by an external party** (e.g. software hack)*	0	90	0	0

Figure 1. An example of a choice task in experiment A

- Experiment B

Instead of presenting the current reference levels, experiment B provided respondents with projections of future reference levels for each attribute. We varied the reference levels shown to the respondents in experiment B. Such variations were created as follows: respondents were told that four experts had given their predictions concerning AV- and CV-fatality numbers, and travel time reductions in 2045. Respondents were randomly assigned an expert (I, II, III, or IV) to see his/her personal prediction in terms of expected fatality numbers, before revealing their preference for a

[6] The original questionnaire was in Dutch.

particular policy package. Table 1 shows the levels embedded in the four treatment variations of experiment B, as well as the benchmark levels used in experiment A. Note that in experiment B, reference levels for AV-fatalities were increased, reference levels for CV-fatalities were decreased (both compared to the current situation). Figure 2 shows an example choice task accompanied by the reference level given by Expert I's estimates.

Table 1. Overview of the reference levels in experiments A and B

	Experiment A	**Experiment B**			
	Current situation	*Expert I*	*Expert II*	*Expert III*	*Expert IV*
Average reduction in travel time	0%	-10%	0%	-5%	-15%
Fatalities caused by CVs / year	600	400	350	300	260
Fatalities caused by technical failure of the AV / year	0	80	160	110	170
Fatalities caused by deliberate misuse of the AV / year	0	50	20	80	60

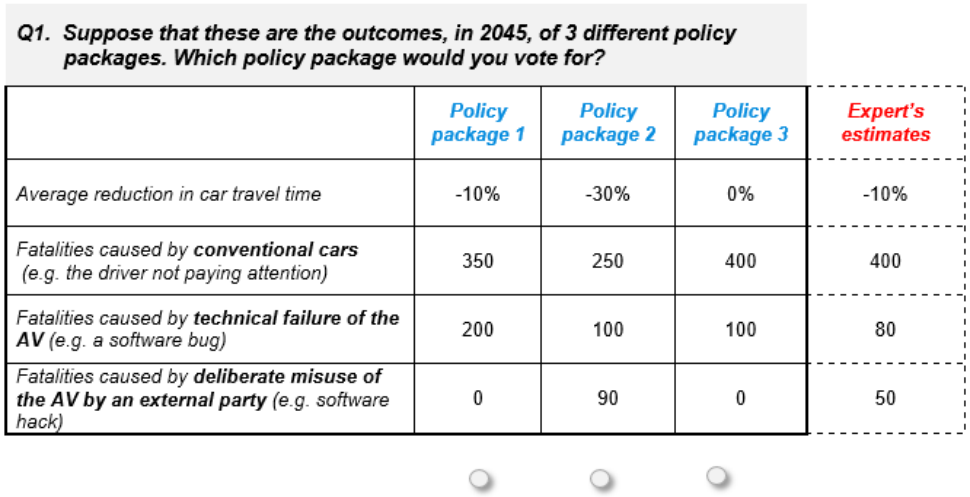


Figure 2. An example of a choice task in experiment B

The scientific aim behind the treatment is that the creation of different (compared to the current situation) reference points allows for the identification of possible reference point effect as discussed in the Introduction. More specifically, we hypothesize that a (partial) reason behind the overweighting of AV-fatalities relates to the simple fact that current levels of AV-related fatalities are extremely low (zero AV-related fatalities), compared to the current level of AV-fatalities.

- Reflection question

After the choice experiment, respondents were also presented with a reflection question. In it, we asked them to evaluate the extent to which they actually considered the reference level presented to them. In experiment A, respondents were asked to indicate, on a five-point Likert scale (ranging from strongly disagree to strongly agree, to what extent they agreed with the following proposition: "I considered the current situation when making choices." , and in experiment B, the

proposition read: "I considered the expert's estimates when making choices." The distribution of answers will be shown in Section 2.3.

Finally, the whole procedure of our experiments is depicted in Figure 3.

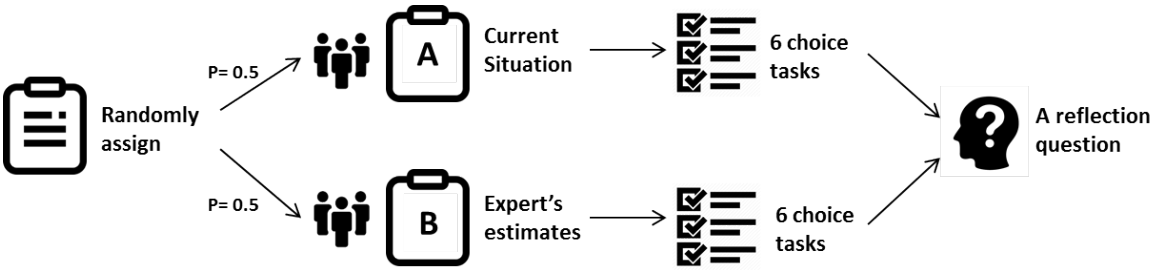


Figure 3. The procedure of the choice experiment

2.3 Data collections

The data collection was conducted during late April and early May, 2018 in the Zuid-Holland province of the Netherlands; specially within the cities of The Hague and Delft. Respondents were approached at random in the vicinity of public parking facilities within an invitation to fill out, on the spot, a paper-pencil survey, or within a flyer containing the URL and QR-code of the survey[7]. A final sample of 412 completed questionnaires, filled out by individuals who were at least 17 years old[8], were obtained via either the paper-pencil (N=232) or online (N=180) version. Each participants was randomly assigned to one of the two experiments, leading to a final sample of 197 individuals for experiment A, and of 214 individuals for experiment B. Note that this random assignment to either the survey without or the one with experimental treatment (i.e., the presence of an artificial reference point) provides a stronger mechanism to identify causal treatment-effects than if we were to ask each individual to make choices in the context of both the surveys with and without treatment. The socio-demographic characteristics of samples A and B are shown in Table 2.

Table 2. Socio-demographic and other relevant characteristics of the sample

Variables		Experiment A	Experiment B
Gender	Male	62.4%	58.1%
	Female	35.0%	39.5%
	Unknown	2.5%	2.4%
Age	17-25	34.5%	41.4%
	26-50	41.1%	30.7%
	51-70	21.4%	22.8%
	Unknown	3%	5.1%
Completed education	Primary school	1.5%	1.8%
	Middle school	13.2%	16.3%
	High school	20.8%	21.4%
	Bachelor	48.2%	49.8%
	Master or doctor	14.7%	10.7%
	Unknown	1.5%	0.0%
Current usage of Advanced Driver Assistance System (ADAS) (e.g. Adaptive cruise control, lane departure warning, automated park assistant)	Yes	44.7%	42.8%
	No	53.8%	54.0%
	Unknown	1.5%	3.2%

[7] In case some respondents indicated that they did not have enough time.
[8] In the Netherlands, citizens are allowed to have a driving license when they are at 17 year old.

Experience with minor accidents	Yes	40.6%	41.4%
	No	58.4%	57.7%
	Unknown	1.0%	1.0%
Experience with severe accidents	Yes	10.7%	9.2%
	No	89.3%	90.3%
	Unknown	0.0%	0.5%

The Chi-square test shows that the samples of experiment A and B did not differ significantly from one another in terms of gender (P=0.37), age (P=0.13), completed education (P=0.37), current usage of ADAS (P=0.70), experience with minor accidents (P=0.83), and experiences with severe accidents (P=0.61), which indicates that two samples are similar along these lines. This implies that differences between experiments in terms of obtained results are due to the experimental treatment, rather than the differences in the samples. After inspecting the socio-demographic variables, we can notice that both samples are slightly skewed towards males and higher educated people. Specifically, about 60% of the participants were males, and about 60% were highly educated in both samples. However, it should be noted here that no attempt was made to arrive at a representative sample from the Dutch population. The reason for this is that our research aims to search for the first empirical evidence for the overweighting of AV-fatalities, compared to CV-fatalities, which could pave the way for more elaborate explorations and confirmatory studies in different regions and countries. Furthermore, we also checked if the socio-demographic factors have an impact on the degree of overweighting, and found no significant results in that regard.

The distribution of the reflection question is shown in Table 3. As expected, respondents to experiment A turned to be more inclined to consider the offered reference point than respondents to experiment B. To respondents, the reference points in experiment B were projections far into the future by an unknown and randomly selected expert, in contrast to reliable information about recent and current levels in experiment A. In next section, we will explore if the level of consideration of reference points has an effect on the overweighting of AV-fatalities.

Table 3. To what extent did participants consider the reference points provided to them?

	Experiment A: ***Reference point is Current situation***	**Experiment B:** ***Reference point is Expert's estimate***
Strongly disagree	6.6%	15.3%
Disagree	20.3%	25.6%
Neutral	18.8%	34.0%
Agree	45.7%	23.3%
Strongly agree	8.6%	1.9%

3. Modelling methodology and estimation results

3.1 Modelling methodology

We develop two models to analyze the choices made in experiments A and B. The first model, which is a baseline model, is used to examine whether, and the extent to which, an AV-fatality is overweighted, compared to a CV-fatality. It is estimated on the data of experiment A and B separately. The second model, called the reference-point model, is applied to explore the effect of reference levels on the (over-)weighting of AV- and CV-fatalities. In order to involve all reference points (i.e., current and artificial reference points), the second model is estimated on the pooled data of experiments A and B.

- Baseline model

The baseline model is a simple Logit-model[9] with the following specification for the systematic utility:

$$V_i = \beta_{CF} CF_i + (1 + \lambda_{AFT})\beta_{CF} AFT_i + (1 + \lambda_{AFM})\beta_{CF} AFM_i + \beta_{TR} TR_i. \quad (1)$$

Here, i denotes the alternative (note that alternatives were unlabeled in the experiments). Parameters β_{CF} and β_{TR} refer to fatalities caused by CVs (CF) and average reduction of car travel time (TR) respectively. To facilitate the interpretation of the differences in weighting between (the two types of) AV-fatalities and CV-fatalities, we write coefficients for fatalities caused by AVs as $(1 + \lambda_{AFT})\beta_{CF}$ and $(1 + \lambda_{AFM})\beta_{CF}$ respectively. Here, parameter λ_{AFT} gives the degree to which an AV-fatality caused by technical failure (AFT) is overweighted, compared to a CV-fatality; and parameter λ_{AFM} gives the degree to which an AV-fatality caused by deliberate misuse (AFM) is overweighted, compared to a CV-fatality. If a λ parameter is estimated to be significantly different from zero, this indicates that there is indeed a difference between AV- and CV-fatalities, in terms of the extent to which weight they receive from respondents. By estimating this model on choice data from experiment A, we can examine the overweighting effect of AV-fatalities, compared to CV-fatalities. Furthermore, by comparing estimation results between experiment A and B, we can get a first idea of whether the treatment of artificial reference points embedded in experiment B have had a downward effect on the degree of the overweighting, or not.

Based on the model described in (1) and estimation results of Overakker (2017), we can derive the following hypotheses; we first focus on experiment A:

1. $\lambda_{AFT} > 0$; that is, fatalities caused by technical failure of the AV are overweighted, compared to fatalities caused by CVs.
2. $\lambda_{AFM} > 0$; that is, fatalities caused by deliberate misuse of the AV are overweighted, compared to fatalities caused by CVs.
3. $\lambda_{AFM} > \lambda_{AFT}$; that is, fatalities caused by deliberate misuse of the AV are overweighted, compared to fatalities caused by technical failure of the AV.

If we find support for these three hypotheses, it implies a qualitative (proxy-) replication of results of the results obtained in Overakker (2017). The size of the overweighting found in our experiments could however still be different.

Moving to a comparison between experiment A and B, the following additional hypotheses[10] are considered:

4. $\lambda_{AFT\_A} > \lambda_{AFT\_B}$; that is, the estimate of λ_{AFT} based on data from experiment A is larger than the corresponding estimate based on data from experiment B, indicating that the overweighting of AV-fatalities caused by technical failure is largest, when the reference point is the current situation (i.e., 0).
5. $\lambda_{AFM\_A} > \lambda_{AFM\_B}$; that is, the estimate of λ_{AFM} based on data from experiment A is larger than the corresponding estimate based on data from experiment B, indicating that the overweighting of AV-fatalities caused by deliberate misuse is largest, when the reference point is the current situation (i.e., 0).

Together these two hypotheses, if we find support for them, suggest that at least part of the overweighting of AV-fatalities should be attributed not to the intrinsic differences between AVs

[9] A series of Mixed logit models was estimated as well, to take into account the panel structure of the data and to allow for heterogeneity within the sample in terms of the weight for fatalities and travel time. These models gave qualitatively similar results compared to the Logit model results reported here, and warrant the same conclusions. For clarity of exposition, we limit ourselves to discussing the Logit model outcomes in this paper.

[10] Note that we compare the parameters which capture differences in weighting of fatalities (i.e., λ_{AFT} and λ_{AFM}) between two experiments, rather than the coefficients of fatalities themselves.

and CVs, but to the simple fact that AV-fatality levels are currently extremely low, making any additional fatality stand out.

- Reference-point model

To further explore the effect of reference points on the weighting of CV- and AV-fatalities, we propose another Logit model based on the following of specification of the systematic utility, which is estimated on the pooled data of experiments A and B:

$$V_i = (\beta_{CF} + \gamma_{CF}[600 - R_{CF}])CF_i + (\beta_{AFT} + \gamma_{AFT}R_{AFT})AFT_i + (\beta_{AFM} + \gamma_{AFM}R_{AFM})AFM_i + \beta_{TR}TR_i \quad (2)$$

Here, R_{CF} gives the prevailing reference point for CV-fatalities; R_{AFT} and R_{AFM} give the prevailing reference points for AV-fatalities caused by technical failure, and caused by deliberate misuse, respectively[11]. Parameter γ_{CF} represents the effect of a one-unit change in the prevailing reference point for CV-fatalities, which is added to the reference-free weight (β_{CF}) associated with a CV-fatality. Likewise, parameter γ_{AFT} represents the effect of a one-unit change in the prevailing reference point for AV-fatalities cause by technical failure, which is added to the reference-free weight (β_{AFT}) associated with such an AV-fatality; and parameter γ_{AFM} represents the effect of a one-unit change in the prevailing reference point for AV-fatalities caused by deliberate misuse, which is added to the reference-free weight (β_{AFM}) associated with such an AV-fatality. It is important to note here, that in the current situation (i.e., as in experiment A), $R_{CF} = 600$ and $R_{AFT} = R_{AFM} = 0$; so if we substitute the reference levels of the current situation into Equation (2), the equation reduces to the following, simplified utility specification: $V_i = \beta_{CF}CF_i + \beta_{AFT}AFT_i + \beta_{AFM}AFM_i + \beta_{TR}TR_i$.

Based on the model described in (2), we can derive the following additional hypotheses:

6. $\gamma_{CF} < 0$; that is, fatalities caused by CVs receive more (negative) weight, as the reference level becomes lower than the current level of 600 CV-fatalities.
7. $\gamma_{AFT} > 0$ and $\gamma_{AFM} > 0$; that is, as the reference point for AV-fatalities increases, the (negative) weight assigned to such fatalities become smaller.

If we find empirical support for these two hypotheses, it would reinforce the idea that the reference effect plays a role in explaining the higher weight attached to AV-fatalities, compared to CV-fatalities; and more generally, that reference points co-determine the weight attached to additional traffic fatalities of different types.

3.2 Model estimation results

We start by estimating the baseline model with systematic utility defined as in Equation (1), on both the data obtained through experiment A and the data obtained through experiment B. Table 4 presents model estimation results.

Table 4. Estimation results of the base model (Equation 1)

		Experiment A			**Experiment B**		
		Coeff.	***Std. err***	***t-value***	***Coeff.***	***Std. err***	***t-value***
Average reduction in car travel time	β_{TR}	0.032	0.004	7.59	0.016	0.004	4.07
Fatalities caused by conventional cars	β_{CF}	-0.009	0.001	-14.51	-0.008	0.001	-14.09

[11] Note that since in this study we are not interested in reference point effects on travel time sensitivity, we include the travel time attribute in our model in a reference-free fashion.

Fatalities caused by AV technical failure	λ_{AFT}	0.541	0.118	4.60	0.122	0.113	1.08*
Fatalities caused by AV deliberate misuse	λ_{AFM}	1.220	0.170	7.19	0.706	0.164	4.32
Number of observations		1182 (i.e., 197×6)			1290 (i.e., 215×6)		
Null LL		-1299			-1417		
Final LL		-1123			-1274		
Rho-square		0.133			0.101		

* not significantly different from 0 at the 95% confidence level (two-tailed test).

Starting with the column representing experiment A, it is easily seen that support is found for all first three hypotheses. That is, the fatalities caused by AVs are weighted more than fatalities caused by CVs, and the fatalities caused by deliberate misuse of the AV receive higher weight than fatalities caused by technical failure (i.e., $2.2 > 1.5$). The asymptotic t-value for the difference between parameters λ_{AFT} and λ_{AFM} equals 3.28 , meaning that these two types of AV-fatalities are weighted differently by participants. If we compare our estimation results with the ones by Overakker (2017), we can find that the degrees of overweighting are smaller: 1.5 as compared to 4 (technical failure) and 2.2 as compared to 5.5 (deliberate misuse). Such differences are expected, given that various changes in experimental designs and also different samples. Our results can, however, still be considered to be in line with those reported by Overakker (2017).

Comparing the estimation results between experiment A and B, we find clear support for hypotheses 4 and 5 (asymptotic t-values for the differences are 2.56 and 2.17 respectively); that is, we find that the degree of overweighting is substantially reduced when reference points for the AV are no longer zero fatalities (but higher), and the reference point for CV is no longer 600 fatalities (but lower). While there is still significant overweighting for deliberate misuse, this is no longer the case for fatalities caused by technical failure of the AV. This provides a strong suggestion that at least part of the overweighting of AV-fatalities compared to CV-fatalities is due to currently very low reference levels for AV-fatalities and a very high level for CV-fatalities.

The reference-point model based on systematic utility specified as Equation (2) is estimated on the pooled data of experiments A and B, the estimation results are presented in Table 5. It can be seen, that no support is found for hypothesis 6; that is, we do not find a lower—than the current level of 600—reference point for CV-fatalities leads to a significantly higher weight carried by per additional CV-fatality. Although the signs of γ_{AFT} and γ_{AFM} are as expected—meaning that as the reference point for AV-fatalities increases from zero, the (negative) weight assigned to such fatalities becomes smaller, these effects are not significant at 95% levels of significance.

Table 5. Estimation results of the reference-point model

	Experiment A + Experiment B		
	Coeff.	***Std. err***	***t-value***
β_{TR}	0.0231	0.0028	8.17
β_{CF}	-0.0082	0.0006	-14.44
γ_{CF} (per $R_{CF}/100$)	0.0002	0.0003	0.61*
β_{AFT}	-0.0119	0.0008	-14.72
γ_{AFT} (per $R_{AFT}/100$)	0.0010	0.0007	1.59*

β_{AFM}	-0.0172	0.0013	-13.83
γ_{AFM} (per $R_{AFM}/100$)	0.0026	0.0023	1.16*
Number of observations	2472 (i.e.,197×6+215×6)		
Null LL	-2716		
Final LL	-2401		
Rho-square	0.116		

* not significantly different from 0 at the 95% level (two-tailed test).

However, it should be noted here that, as highlighted in Section 2 (Table 3), a substantial share of respondents (40.9%) in experiment B indicated that they did not actually consider the artificial reference points when making choice. It is interesting to separately analyze the subsample who indicated that they did consider the reference point. When excluding those participants who indicated that they either disagreed or strongly disagreed with the statement "I considered the expert's estimates when making choices", a different picture arises, as can be seen in Table 6.

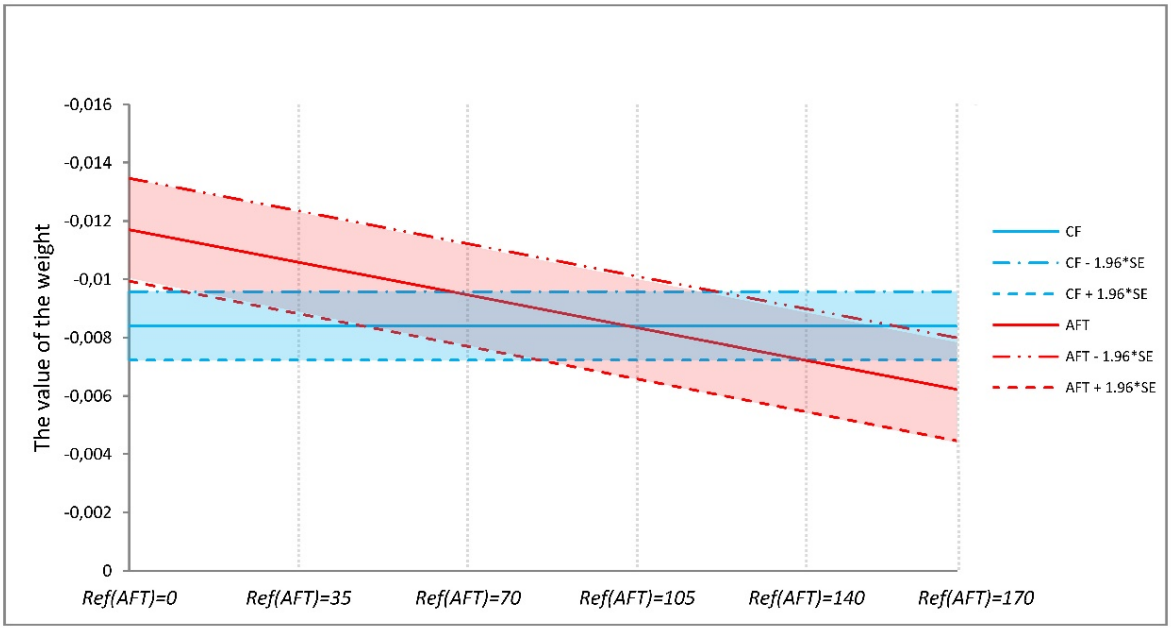
Table 6. Estimation results of the reference-point model, subsample of participants who did not express (strong) disagreement with the following proposition: *"I considered the expert's estimates when making choices"*

	Experiment A + Experiment B (subsample)		
	Coeff.	***Std. err***	***t-value***
β_{TR}	0.0211	0.0032	6.65
β_{CF}	-0.0084	0.0006	-14.22
γ_{CF} (per $R_{CF}/100$)	0.0000	0.0003	0.03 *
β_{AFT}	-0.0117	0.0009	-13.61
γ_{AFT} (per $R_{AFT}/100$)	0.0032	0.0008	4.13
β_{AFM}	-0.0172	0.0013	-12.87
γ_{AFM} (per $R_{AFM}/100$)	0.0113	0.0028	4.10
Number of observations	1944 (i.e.,197×6+127×6)		
Null LL	-2136		
Final LL	-1896		
Rho-square	0.112		

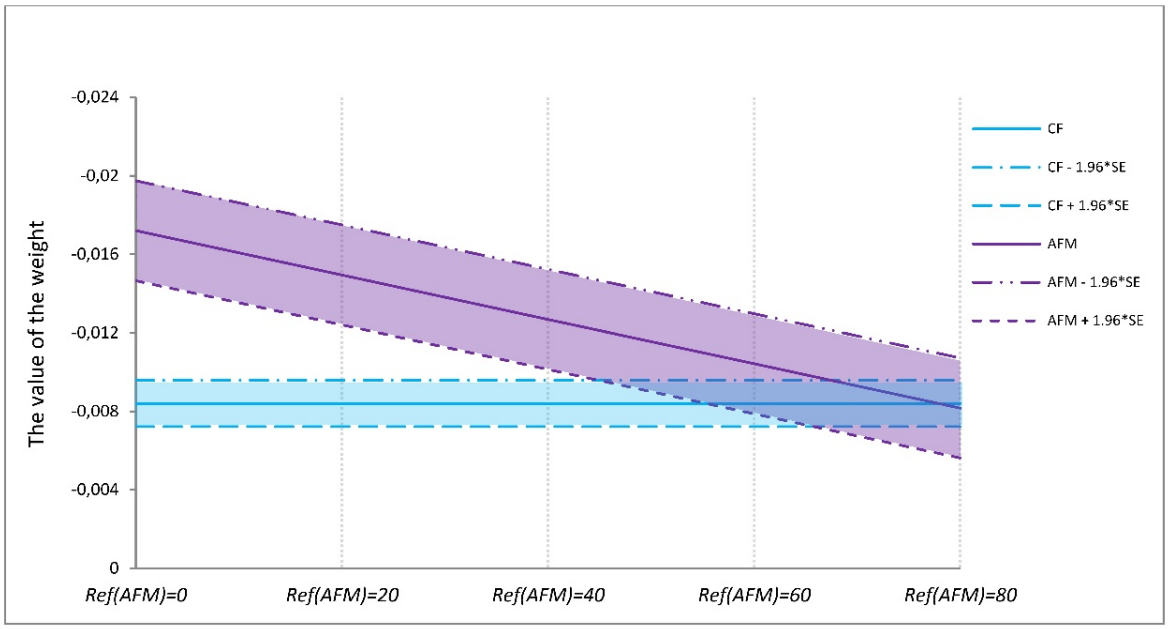
* not significantly different from 0 at the 95% level (two-tailed test)

Table 6 shows that, for the subsample of respondents who indicated that they considered the artificial reference points in experiment B, we clearly find support for the hypothesis 7, that is increases in the reference point concerning AV-fatalities lead to a decline in the weight associated with such fatalities. This effect is strongest for fatalities caused by AV-deliberate misuse, but also exists for fatalities caused by AV-technical failure. This finding is further illustrated in Figure 4, where we plot the trends in fatality-weights including their 95% confidence intervals. It is clearly seen that as AV-fatalities move away from the current reference point of zero, at some point there is no overweighting left, relative to CV-fatalities. Furthermore, we notice that parameter γ_{CF} is not significant, suggesting that decreasing the reference level does not influence the weighting of CV-fatalities. One possible explanation is that people's perceptions and beliefs regarding CV-fatalities

are relatively stable compared to their perceptions and beliefs regarding new types of fatalities, e.g., AV-fatalities; as a result, changing reference levels does not affect how they view and weigh CV-fatalities.



(a) AFT



(b) AFM

Figure 4. The trends of the weight associated with AV-fatalities for increasing values of the corresponding reference points (inverted scales)

4. Conclusions and discussions

By analysing the choices made by respondents in our first SC experiment, we found that fatalities caused by AVs received more weight than fatalities caused by human drivers in CVs. The

difference in weighting equalled 54% for AV-fatalities caused by technical failure of the AV (e.g., software bugs), and 122% for AV-fatalities caused by deliberate misuse of the AV (e.g., software hack). These degrees of overweighting are somewhat smaller than, but are still within the same order of magnitude as, the results in a previous SC experiment (Overakker, 2017). In the second SC experiment, we explored a specific potential reason for this overweighting of AV-fatalities. Specifically, we hypothesized that the current levels of AV-fatalities are so low (i.e., zero) that any additional AV-fatality carries more weight, this having little to do with intrinsic differences between AV- and CV-fatalities in public perception. As hypothesized, we found that by artificially increasing the reference levels of AV-fatalities, we were able to substantially reduce respondents' overweighting of AV-fatalities caused by deliberate misuse, and even eliminate the overweighting of AV-fatalities caused by technical failure.

Delving into the reasons behind the reference level effect, we can find compelling explanations in social science literature. First, *loss aversion* in prospect theory (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991) suggests that losses are weighted more heavily than equivalent gains, in this case implying that the increase (loss) in AV-fatalities (relative to the current reference point of zero) will receive more weight than a corresponding decrease (gain) in CV-fatalities. Second, *probability weighting* in prospect theory may also offer a partial explanation: given the currently very low probability of having a fatal accident with an AV, the perceptions of this risk are inflated due to the tendency to overweight small probabilities. Third, and probably most importantly, the so-called Weber effect (Gescheider, 2013; Weber, 1834) suggests that a change relative to a low baseline is perceived as bigger than the same—in an absolute sense—change compared to a higher base-level. In our case, this implies that a change from, for example, zero to ten AV-fatalities weighs more heavily than a change from 600 to 590 or to 610 CV-fatalities.

What implications does this study have for the public debate surrounding the topic of safety issues of AVs, and for policy making in this domain? A first implication of our results, is that the safety paradox affecting AVs—i.e., the notion that while AVs are expected to bring great traffic safety benefits, problematic safety perceptions may delay or even prohibit their introduction—seems to be less salient than many experts have suggested in the public debate. For example, our results suggest that cutting the number of annual fatalities in half by implementing AVs in the Netherlands would already be considered acceptable by the Dutch citizens. This is a much smaller decrease than what is often heard in policy and public debates. Furthermore, our findings suggest that as the number of AV-fatalities increases—as will inevitably be the case, once they are more and more allowed to drive on real roads—the extent to which they will be overweighted compared to CV-fatalities will decrease (ceteris paribus). This implies that, somewhat ironically, the occurrence of accidents involving AVs will help redress the above-mentioned safety paradox surrounding AVs. In combination, our findings suggests that once AVs can save at least half of the number of lives lost annually in traffic accidents, there should be no reason to fear that safety perception issues among the general public will backfire to the extent that AV-acceptance becomes highly problematic. In that sense, we concur with Kalra and Groves (2017) that "the perfect should not be the enemy of the good": once AVs have become considerably safer than CVs, they should be allowed on the road as soon as possible, to speed up their learning process (making them even safer) and to save, during this process, lives that would otherwise have been lost in accidents involving CVs.

Clearly, these conclusions and implications are based on a study which has its limitations: first, we used SC-experiments involving participants making hypothetical choices, based on fatality-statistics. Although we believe that this type of experiment is well suited to answer the type of questions posed in this paper—that is, to infer weight from choices, rather than simply asking people for their weight directly—the hypothetical nature of our experiment should be kept in mind when interpreting results. Likewise, although accident statistics (like the ones used in our experiments) tend to play an important role in public debates about road safety, it is likely that other, non-numeric, discussions of AV-safety issues in the media and public debate may affect the

safety debate surrounding AVs in ways that go beyond the scope of the conclusions that may be drawn from our study. Furthermore, it should once again be brought to attention that our sample is a relatively small convenience sample recruited in a confined urban area within the Netherlands. Although we did not find evidence for any effect of socio-demographic variables on our main results, it is to be expected that the degree of overweighting of AV-fatalities—compared to CV-fatalities—will be country and (car-) culture specific. Besides, we applied two means of recruiting respondents, immediate responses (i.e., paper-pencil survey) and online responses. Although these two approaches combined are often used in SC experiments, it may bring unknown bias to results as immediate responses usually have less dwell time to think about questions. Last but not least, in this study, we merely look at the reference effect on the overweighting of AV-fatalities, but there can be other possible explanations. An intriguing one from social psychology may also provide an explanation: the general public holds new technologies to higher standards than traditional ones (Fischhoff et al., 1978; Otway & Von Winterfeldt, 1982). Nevertheless, we consider our small-scale study to provide a potential stepping stone for future studies that involve representative samples from different countries where AVs will hit the road in the near future.

Acknowledgements

The first author gratefully acknowledges financial support for her PhD study from the China Scholarship Council (No. 201608310106). The third author wishes to thank the European Research Council for financial support (ERC Consolidator grant BEHAVE—724431).

References

AAA. (2017). Americans feel unsafe sharing the road with fully self-driving cars. Retrieved from https://newsroom.aaa.com/2017/03/americans-feel-unsafe-sharing-road-fully-self-driving-cars/

Anderson, J. M., Nidhi, K., Karlyn, D. S., Paul, S., Constantine, S., & Tobi, A. O. (2016). Autonomous vehicle technology: A guid for policymakers. Retrieved from https://www.rand.org/pubs/research_reports/RR443-2.html

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

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

de Almeida Correia, G. H., Looff, E., van Cranenburgh, S., Snelder, M., & van Arem, B. (2019). On the impact of vehicle automation on the value of travel time while performing work and leisure activities in a car: Theoretical insights and results from a stated preference survey. *Transportation Research Part A: Policy Practice, 119*, 359-382. doi:https://doi.org/10.1016/j.tra.2018.11.016

Economist, T. (2018). How do you define "safe driving" in terms a machine can understand? Writing the robotic rules of the road. Retrieved from https://www.economist.com/science-and-technology/2018/05/10/how-do-you-define-safe-driving-in-terms-a-machine-can-understand

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. *Transportation Research Part A: Policy Practice, 77*, 167-181. doi:https://doi.org/10.1016/j.tra.2015.04.003

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy sciences, 9*(2), 127-152. doi:https://doi.org/10.1007/BF00143739

Gescheider, G. A. (2013). *Psychophysics: the fundamentals*: Psychology Press.

Greenblatt, J. B., & Saxena, S. (2015). Autonomous taxis could greatly reduce greenhouse-gas emissions of US light-duty vehicles. *Nature Climate Changes, 5*(9), 860-863. doi:https://www.nature.com/articles/nclimate2685#supplementary-information

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263-292.

Kalra, N., & Groves, D. G. (2017). *The enemy of good: estimating the cost of waiting for nearly perfect automated* Vehicles. Retrieved from https://www.rand.org/pubs/research_reports/RR2150.html

Kyriakidis, M., Happee, R., de Winter, J. C., & behaviour. (2015). Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation research part F: traffic psychology, 32*, 127-140. doi:https://doi.org/10.1016/j.trf.2015.04.014

Litman, T. (2019). *Autonomous vehicle implementation predictions*. Retrieved from https://www.vtpi.org/avip.pdf

Liu, P., Yang, R., & Xu, Z. (2019). How safe is safe enough for self-driving vehicles? *Risk analysis, 39*(2), 315-325. doi:https://doi.org/10.1111/risa.13116

Mervis, J. (2017). Are we going too fast on driverless cars? *Science*. Retrieved from http://www.sciencemag.org/news/2017/12/are-we-going-too-fast-driverless-cars

Nees, M. A. (2019). Safer than the average human driver (who is less safe than me)? Examining a popular safety benchmark for self-driving cars. *Journal of Safety Research, 69*, 61-68. doi:https://doi.org/10.1016/j.jsr.2019.02.002

Nieuwenhuijsen, J., de Almeida Correia, G. H., Milakis, D., van Arem, B., & van Daalen, E. (2018). Towards a quantitative method to analyze the long-term innovation diffusion of automated vehicles technology using system dynamics. *Transportation Research Part C: Emerging Technologies, 86*, 300-327. doi:https://doi.org/10.1016/j.trc.2017.11.016

Otway, H. J., & Von Winterfeldt, D. (1982). Beyond acceptable risk: On the social acceptability of technologies. *Policy sciences, 14*(3), 247-256. doi:https://doi.org/10.1007/BF00136399

Overakker, B. (2017). *The Social Acceptance of Automated Driving Systems: Safety Aspects: A contribution to* responsible *innovation by using a referendum format, discrete choice model experiment to measure the social acceptance of ADS by Dutch citizens with corresponding heterogeneity*. (Master), Delft University of Technology, Retrieved from https://repository.tudelft.nl/islandora/object/uuid:31159f82-e33d-4124-a5a2-2b605974870e

Piao, J., McDonald, M., Hounsell, N., Graindorge, M., Graindorge, T., & Malhene, N. (2016). Public views towards implementation of automated vehicles in urban areas. *Transportation research procedia, 14*(0), 2168-2177. doi:https://doi.org/10.1016/j.trpro.2016.05.232

Rose, J. M., & Bliemer, M. C. (2009). Constructing efficient stated choice experimental designs. *Transport* Reviews, *29*(5), 587-617. doi:https://doi.org/10.1080/01441640902827623

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour, 1*(10), 694-696. doi:https://doi.org/10.1038/s41562-017-0202-6

Shladover, S. E., Su, D., & Lu, X.-Y. (2012). Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transportation Research Record, 2324*(1), 63-70. doi:https://doi.org/10.3141/2324-08

Simonite, T. (2013). *Data shows Google's robot cars are smoother, safer drivers than you or I*. Retrieved from https://www.technologyreview.com/s/520746/data-shows-googles-robot-cars-are-smoother-safer-drivers-than-you-or-i/

Singh, S. (2015). *Critical reasons for crashes investigated in the national motor vehicle crash causation survey*. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812506

Sparrow, R., & Howard, M. (2017). When human beings are like drunk robots: Driverless vehicles, ethics, and the future of transport. *Transportation Research Part C: Emerging Technologies, 80*, 206-215. doi:https://doi.org/10.1016/j.trc.2017.04.014

SWOV. (2019). *Road deaths in the Netherlands*. Retrieved from https://www.swov.nl/en/facts-figures/factsheet/road-deaths-netherlands

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics, 106*(4), 1039-1061. doi:https://doi.org/10.2307/2937956

Weber, E. H. (1834). *De pulsu, resorptione, auditu et tactu: annotationes anatomicae et physiologicae, auctore*: prostat apud CF Koehler.