Autonomous Vehicles: How Perspective-Taking Accessibility Alters Moral Judgments and Consumer Purchasing Behavior

Rose Martin, Petko Kusev, and Paul van Schaik

Author Note

Rose Martin, Huddersfield Business School, and The Behavioural Research Centre, The University of Huddersfield; Petko Kusev, Huddersfield Business School, and The Behavioural Research Centre, The University of Huddersfield; Paul van Schaik, Department of Psychology, Teesside University. Rose Martin was supported by The University of Huddersfield Early Career Scholarship. We are grateful to Hristina Kuseva and Kath Lopez for their comments on the manuscript draft.

Correspondence concerning this article should be addressed to Petko Kusev, The University of Huddersfield, Huddersfield Business School, Huddersfield, HD1 3DH, United Kingdom, e-mail: p.kusev@hud.ac.uk, Phone: +44 (0)1484473290; Fax: +44(0)1484 516151.

Abstract

In preparation for unavoidable collisions, autonomous vehicle (AV) manufacturers could program their cars with utiliarian ethical algorithms that maximize the number of lives saved during a crash. However, recent research employing hypothetical AV crash scenarios reveals that people are not willing to purchase a utilitarian AV despite judging them to be morally appropriate (Bonnefon et al., 2016). This important result, indicating evidence for a social dilemma, has not yet been psychologically explored by behavioral scientists. In order to address the psychological underpinnings of this phenomenon, we developed and tested a novel theoretical proposal – *perspective-taking accessibility* (PT accessibility). Accordingly, we established that providing participants with access to both situational perspectives (AV buyers can be passengers or pedestrians) in crash scenarios, eliminated the behavioral inconsistency between their utilitarian judgments of moral appropriateness and non-utilitarian purchasing behavior. Moreover, our full PT accessibility induced respondents' utilitarian prosocial judgments and purchasing behavior (Experiments 1a and 1b) and consistent utilitarian preferences across judgment tasks (Experiment 2). Crucially, with full PT accessibility, participants' utilitarian purchasing behavior as well as their willingness to buy and ride utilitarian AVs were informed by their utilitarian moral judgments. Full PT accessibility provides the participants with even odds of being a pedestrian or passenger in crash scenarios, and thus impartiality. It could be argued that full PT accessibility is a new type of 'veil of ignorance', which is not based on purposely induced self-interest and uneven risk options (as in Huang et al., 2019), but rather is based on even odds of being a passenger or pedestrian, and therefore with even 50/50 chance to die/live as passenger or pedestrian. Under these circumstances one can measure utilitarian preferences.

Keywords: perspective-taking, accessibility, social dilemma, moral judgments, purchasing behavior, autonomous vehicles

Autonomous Vehicles: How Perspective-Taking Accessibility Alters Moral Judgments and Consumer Purchasing Behavior

1. Introduction

Some decisions made to improve the safety of one individual, can consequently impede the safety of many others. For instance, purchasing an autonomous vehicle (AV) that puts the passenger's safety first could endanger the lives of other drivers and pedestrians. Recent research regarding autonomous vehicle acceptance revealed that people would prefer to buy a passenger-protective car over a prosocial car that protects majority of people during collisions, despite judging the latter as more moral (Bonnefon et al., 2016). The authors explained these two findings as a social dilemma; people know that the world would be safer if everyone bought a utilitarian AV, but they all have a strong personal incentive to do otherwise (Bonnefon et al., 2016; Kollock, 1998).

This important result has not yet been psychologically explored by behavioral scientists. In order to address the psychological reasons for this social dilemma, we have developed and tested a novel theoretical proposal – *perspective-taking accessibility* (PT accessibility) – where the decision-makers have access to full or partial situational perspectives and behavioral consequences within the same moral scenarios. Moreover, we argue that Bonnefon and colleagues' intriguing and important findings are caused by partial PT accessibility in crash scenarios. We found that presenting participants with both the perspective of the car passenger and the pedestrian in hypothetical crash scenarios (full PT accessibility) eliminates the behavioral inconsistency between their utilitarian judgments of moral appropriateness and non-utilitarian purchasing behavior. Accordingly, consistent with (and informed by) their moral judgments, participants were more willing to buy, ride and spend more money on prosocial cars over passenger-protective vehicles.

1.1. Autonomous vehicles and consumer preferences

The introduction of commercially available autonomous vehicles (AVs) has received widespread multidisciplinary and interdisciplinary attention from researchers in artificial intelligence, engineering, transport, law, philosophy, psychology and business (e.g., Awad et al., 2018; Bigman, & Gray, 2018; Bonnefon et al., 2016; de Melo et al., 2019; Fagnant & Kockelman, 2015; Faulhaber et al., 2018; Goodall, 2014; Maurer et al., 2015; Nyholm & Smids, 2016). One implication of replacing human drivers with automated transport systems is the anticipated reduction in the number of road accidents often caused by drink driving, fatigue, and human error (Fagnant & Kockelman, 2015; Goodall, 2014). Moreover, even in situations where a crash is imminent and unavoidable, AVs can be pre-programmed to make split-second ethical calculations in order to determine harm-minimizing trajectories. However, there are different ways that harm can be minimized in the context of potential collisions involving AVs. For instance, following prosocial utilitarian principles, AVs could be programmed to, where possible, minimize overall harm in order to protect the greatest number of people. In some circumstances, this could mean sacrificing the passenger in order to save a greater number of pedestrians. Alternatively, following a passenger-protective agenda, AVs could be programmed to protect the passenger at all costs (even at the expense of other drivers and pedestrians; Gogoll & Müller, 2017; Nees, 2016).

Recent research into AV acceptance reveals that decision-makers are not willing to purchase the prosocial utilitarian AV that they judged to be the most moral (Bonnefon et al., 2016). In particular, Bonnefon and colleagues presented participants with moral scenarios that involved a perspective-taking task; participants were required to imagine themselves as a passenger inside an AV that is about to crash into 10 pedestrians. Accordingly, the participants then judged whether it was more morally appropriate for the AV to stay on its present course (killing the 10 pedestrians) or swerve to the side of the road and crash into a barrier (killing the passenger). The majority of participants indicated that it was more morally

appropriate for the AV to swerve and kill the passenger. However, when considering purchasing an AV, the participants wanted passenger-protective vehicles rather than prosocial utilitarian models. Bonnefon et al.'s (2016) recent example of social dilemma has novel and significant implications for society, since they paradoxically suggest that regulating for prosocial AVs can result in the rejection of AVs, and in turn increase casualties. Specifically, Bonnefon et al. (2016) note that it may therefore be counterproductive for policymakers to enforce utilitarian regulations on AV ethical algorithms, as "such regulation could substantially delay the adoption of AVs, which means that the lives saved by making AVs utilitarian may be outnumbered by the deaths caused by delaying the adoption of AVs altogether" (p. 1575-1576).

AV dilemmas, unlike traditional moral dilemmas (e.g., the trolley problem; Foot, 1967, Thomson, 1985), engage the participants in a perspective-taking task, where the participants are typically asked to imagine themselves as a passenger inside an AV, and therefore a potential victim of their own moral convictions. However, we have previously argued (see Martin et al., 2017) that although the AV dilemma is a perspective-taking task, the task itself lacks access to the situational perspectives of both passenger and pedestrian. For example, in all variations of Bonnefon et al.'s (2016) AV dilemmas, the participants were only given the perspective of the AV passenger, limiting their situational perspectives. This partial PT accessibility is problematic, considering that by default AV buyers will not only be passengers of AVs but will be pedestrians too (e.g., as soon as they exit their vehicle). We consequently argue that AV buyers should be provided with access to both the passenger and pedestrian perspectives (even odds of being a passenger or pedestrian - as a principle of impartiality), or in other words full *perspective-taking (PT) accessibility* in order to value, buy and ride the utilitarian AVs.

It is plausible that the partial PT accessibility in Bonnefon et al.'s (2016) AV dilemmas (having access to only the passenger's perspective - partial PT accessibility) triggered the behavioral inconsistency between participants' non-utilitarian purchase behavior and their utilitarian judgments of moral appropriateness. Accordingly, presenting the participants with partial PT accessibility (having access to only the passenger's perspective) could have a greater relevance and influence on the behavior with consequences (purchasing behavior – owning the AV) than on the behavior with no consequences (judgments of moral appropriateness - expressing an opinion with no consequences to the participants). For example, having access to only the passenger's perspective inside an AV and a task with behavioral consequences (purchasing behavior – owning the AV) nudged participants towards purchasing passenger-protective vehicles (non-utilitarian purchase behavior; Bonnefon et al., 2016). In contrast, the majority of participants indicated that it was more morally appropriate for the AV to swerve and kill the passenger (utilitarian moral judgment) when the task bore no consequences to the participants (Bonnefon et al., 2016). Accordingly, we argue that presenting participants with AV dilemmas with full PT accessibility could eliminate the behavioral inconsistency between participants' non-utilitarian purchase behavior and their utilitarian judgments of moral appropriateness. We also predict that moral dilemmas with full PT accessibility will boost both participants' utilitarian purchasing behavior and utilitarian judgments of moral appropriateness.

Moreover, in accordance with the 'veil of ignorance' (VOI) principles of self-interest and impartiality – even odds of being each of the people affected by the moral decision (e.g., Harsanyi, 1955, 1975; Huang et al., 2019; Rawls, 1971), Huang et al. (2019) proposed that in VOI reasoning tasks participants' responses will tend to be utilitarian, simply because this maximizes their odds of a good outcome. Huang and colleagues also predicted, and empirically established that the participants aligned their later moral judgments with their

earlier VOI preferences (Huang et al., 2019). Crucially, however, Huang's VOI reasoning tasks are always 'personal' (participant involvement), with induced self-interest manipulation ("Please respond from a purely self-interested perspective"), do not engage the participants in full perspective taking, and, according to the authors, employ even odds of being each of the people affected by the moral decision (see Hare, 2016; Harsanyi, 1955, 1975; Huang et al., 2019). According to Huang et al. (2019), the even odds make the moral decision a matter of 'risk', rather than 'ambiguity'. For example: "If the law requires the autonomous vehicle to swerve in such a situation, you have a 1 out of 10 chance of dying and a 9 out of 10 chance of living. If the law forces the autonomous vehicle to stay on its current path, you have a 1 out of 10 chance of living and a 9 out of 10 chance of dying." (see supplemental materials, Huang et al., 2019). Crucially, in this risky decision task, the swerve option is risk-averse (safe) and utilitarian, and the stay option is risk-seeking (risky) and non-utilitarian, and therefore in reality the odds are not even.

As predicted, Huang and colleagues found that the participants in their experiments made utilitarian decisions and judgments; they chose the risk-averse utilitarian swerve option (9 out of 10 chance of living). This result is also supported by Kusev et al.'s (2016) enhanced utilitarian accessibility findings, where comprehensive information about moral actions and consequences boosted utility maximization in moral choices. However, in our proposal, we provide the participants with full and accessible information regarding both situational perspectives (AV buyers can be passengers or pedestrians) and even odds of being a pedestrian or passenger in crash scenarios (eliminating the opportunity of making self-interest decisions – e.g., by selecting the 'risk averse' option, which happened to be utilitarian), the tasks are not always personal (we have included a stranger involvement condition), and we do not induce self-interest by requiring the participants to answer the questions from a purely self-interested perspective. Moreover, it could be argued that full PT accessibility is a new

type of VOI, which is not based on induced self-interest and uneven risk options (as in Huang et al., 2019), but rather is based on even odds of being a passenger or pedestrian, and therefore with even 50/50 chance to die/live as passenger or pedestrian. Accordingly, we propose that the even odds do not trigger self-interest, risk-related preferences or decision biases and facilitate opportunities for prosocial/utilitarian moral judgments and purchasing behavior.

In previous experimental studies that induce partial perspective-taking (e.g., Batson et al.,1997, 2003; Ruby & Decety, 2004; Stotland, 1969), participants are required to imagine (i) how one would feel in someone else's situation (imagine-self), or (ii) how someone else feels in their situation (imagine-other). Both types of perspective-taking involve taking the perspective of one person in a task or scenario. However, no previous research has explored full PT accessibility, where the decision-maker has access to multiple roles (e.g., themselves or someone else as a passenger and pedestrian) within the same moral scenarios. Accordingly, for the purpose of this article, we explore the influence of PT accessibility (full and partial) on moral judgments and decision-making. In Experiments 1a, 1b and 2, we introduced AV crash scenarios with full PT accessibility, where participants have access to two perspectives; they were invited to consider the possibility of being a passenger and a pedestrian in AV crash scenarios.

Whilst there is a normative expectation that humans are able to make successful inferences and maximize utility in decision-making tasks (e.g., Kahneman, 2003; Kusev et al., 2016, 2020; Tversky & Kahneman, 1973) we argue that PT is not straightforward or an intuitive decision strategy that people engage in. For example, when individuals assess moral scenarios with a focus on only one particular agent (the passenger) they may not even consider that they will inevitably be pedestrians too. Therefore, it is plausible that presenting moral dilemmas with limited PT accessibility, unnecessarily emphasizes the risk of utilitarian AVs and as a consequence disguises the potential benefits of utilitarian AVs (see Shariff et al., 2016).

Accordingly, in three experiments we measured respondents' prosocial moral judgments and purchasing behavior and found that access to full PT accessibility (with even odds) in AV crash scenarios further enhanced their utilitarian moral judgments and utilitarian preference consistency across judgment tasks. Moreover, PT accessibility eliminated the behavioral inconsistency between participants' utilitarian moral judgments and their non-utilitarian purchasing behavior and overall boosted their utilitarian preferences (Experiments 1a and 1b). Crucially, with full PT accessibility, participants' utilitarian purchasing behavior as well as their willingness to buy and ride utilitarian AVs were informed by their utilitarian moral judgments (Experiments 1a, 1b and 2).

2. Experiment 1a

2.1. Method

2.1.1. Participants

In Experiment 1, participants (N = 300; 165 females and 135 males) were recruited via an online recruitment service, and were rewarded £1 for their participation. The mean age of the participants was 52 (SD = 14.93). All participants were treated in accordance with the code of ethics of the World Medical Association (WMA).

For statistical testing, we used a significance level of .05. Although we did not assume an effect size, we wanted to ensure that our sample size would allow us to detect a large effect size (f= .40 by convention; Cohen, 1988) of the independent-measures effects of *type of involvement*, *type of PT accessibility* and their interaction. We ran the experiment for 14 days to ensure that data collection from a sufficiently large sample would achieve a statistical power of at least .95. According to the retrospective power analysis, the achieved sample size (N = 300) produced a power of 1.00 which was sufficient to achieve our target.

2.1.2. Experimental design

A 2x2 independent measures design was employed. The first independent variable, *type of involvement*, had two levels including: (i) participant involvement (where the participant is described as an agent in the scenario) and (ii) stranger involvement (where a stranger is described as an agent in the scenario). This independent variable was equivalent to Batson et al.'s (1997) imagine-self (participant involvement) and imagine-other (stranger involvement) perspective-taking and was also employed by Bonnefon et al. (2016).

The second independent variable, *type of PT accessibility*, also had two levels including: (i) partial PT accessibility (where the agent is described as being inside an AV in the scenario) and (ii) full PT accessibility (where it is made clear that the agent could potentially be inside the AV or part a group of pedestrians in the scenario).

The first dependent variable was *judgments of moral appropriateness* where participants were required to judge on a 10-point rating scale how morally appropriate they believed each AV was (higher numerical ratings denoting a higher judgment of moral appropriateness). It is important to note that Bonnefon and colleagues' (2016) method (placing swerve and stay AVs on a single unipolar scale) did not establish whether each respondent is overall utilitarian (or not) when judging both moral situations (swerve and stay judgments). Moreover, they were also unable to establish the weight/magnitude of utilitarian/non-utilitarian judgments. In order to avoid this pitfall, we used separate moral appropriateness scales for swerve and stay judgments. Therefore, participants' overall judgments of moral appropriateness were computed as a utilitarian weight – the difference between participants' judgments of moral appropriateness for utilitarian swerve and non-utilitarian stay AVs. Specifically, we subtracted the judgments for non-utilitarian stay AVs from judgments of utilitarian swerve AVs in order to generate a utilitarian weight. Therefore, positive and high difference (utilitarian weight) indicated utilitarian judgments.

Similar to the point allocation system employed in Bonnefon et al. (2016), the second dependent variable was *purchasing value,* where participants were given a budget of £50,000 to distribute between the two AV options. Accordingly, purchasing value was calculated as the difference between the amount of money participants are willing to spend on utilitarian swerve and non-utilitarian stay AVs (where value for non-utilitarian stay AVs was subtracted from the value for utilitarian swerve AVs).

2.1.3. Materials and procedure

Participants took part in an online study where they were presented with a scenario involving an AV with 1 passenger inside that is about to crash into a group of 10 pedestrians. The scenario description and visual stimuli were dependent on the condition in which the participant was randomly allocated to. For example, participants who were allocated to the participant involvement and partial PT accessibility condition received the following scenario (adapted from Bonnefon et al. 2016; see also Figure 1A and all moral scenarios in Supplemental Materials):

"You are the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Suddenly, 10 pedestrians appear ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing you but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians, but you will be unharmed."

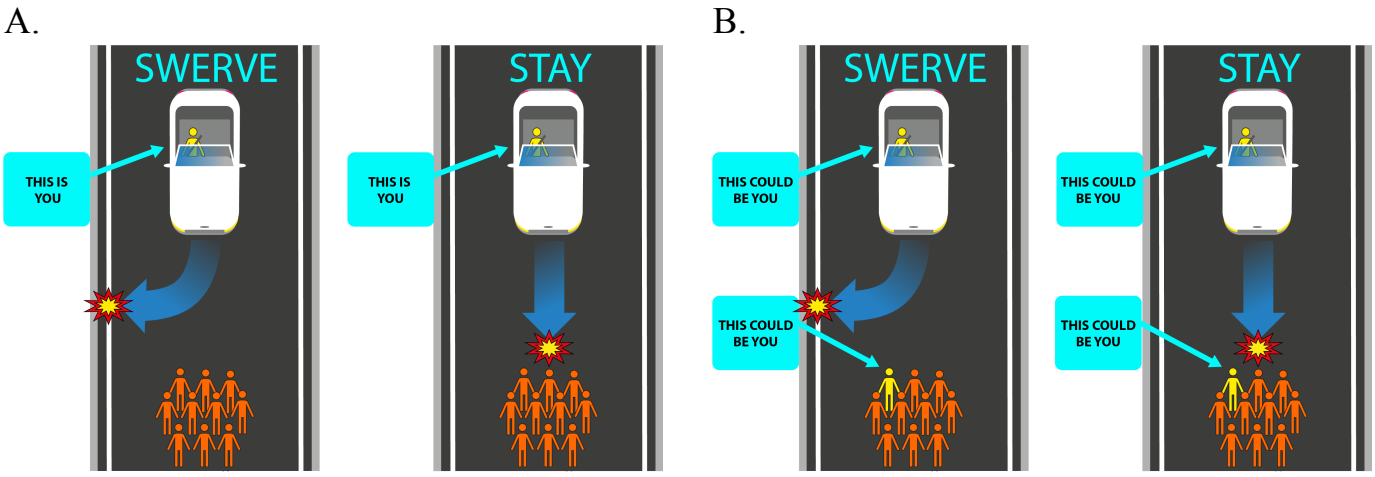
Furthermore, participants allocated to the participant involvement and full PT accessibility condition would have received the following scenario (see also Figure 1B):

"You could be the sole passenger in an autonomous self-driving vehicle travelling at the speed limit down a main road. Or you could be one of the 10 pedestrians that have appeared ahead, in the direct path of the car. The car could be programmed to either **SWERVE** off to the side of the road, where it will impact a barrier, killing the passenger (that could be you)

but leaving the 10 pedestrians unharmed or **STAY** on its current path where it will kill the 10 pedestrians (that could include you), but the passenger will be unharmed."

Figure 1

Visual Stimuli Presented to Participants in the Participant Involvement Condition



Note. Panel A represents partial PT accessibility and Panel B represents full PT accessibility (see Figure S1 in Supplemental Materials for all original stimuli used in Experiments 1a, 1b and 2).

Participants who were allocated to the stranger involvement condition read an identical scenario about a character called Sam (with visual stimuli referring to Sam). After reading the scenario, participants were required to make the following judgments (presented one at a time in a randomized order):

Judge the moral appropriateness of programming a car to swerve;

Judge the moral appropriateness of programming a car to stay; and

Using a budget of £50,000, please indicate how much you would pay for the following autonomous self-driving cars (the entire £50,000 budget must be spent): a car that is programmed to swerve and a car that is programmed to stay.

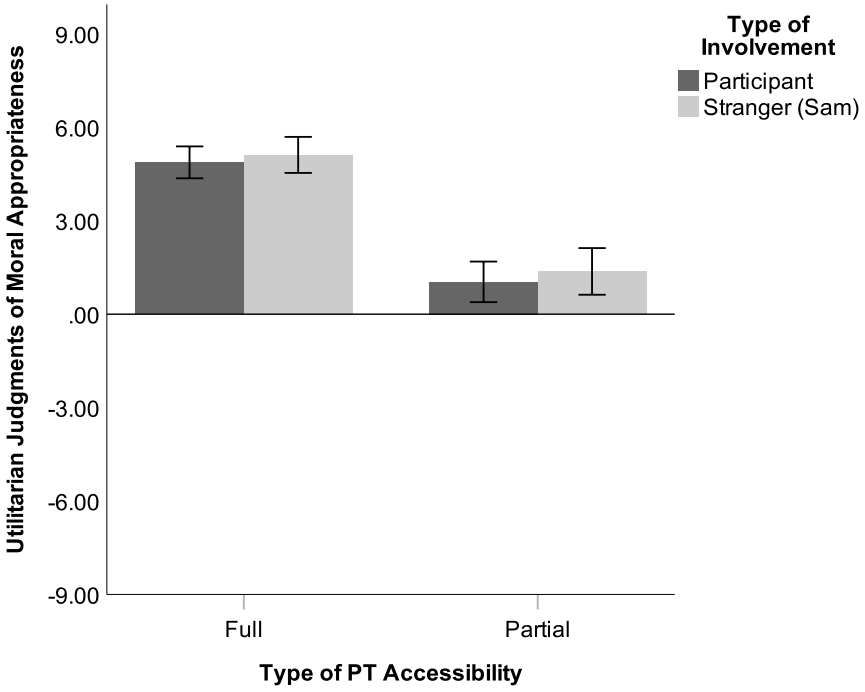
2.2. Results and discussion

2.2.1. Judgments of moral appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on participants' judgments of moral appropriateness. The results revealed that type of PT accessibility significantly influenced respondents' judgments of moral appropriateness $F(1, 296) = 143.90$, $p < .001$, $\eta_p^2 = .33$. Specifically, participants were more utilitarian in their moral judgments with full PT accessibility ($M = 5.00$; $SD = 2.37$) than with partial PT accessibility ($M = 1.21$; $SD = 3.05$). However, the results revealed that main effect of type of involvement ($F < 1$), as well as the two-way interaction of type of involvement by type of PT accessibility ($F < 1$) on participants' judgments of moral appropriateness were not statistically significant (see Figure 2).

Figure 2

Participants' Utilitarian Judgments of Moral Appropriateness in Experiment 1a



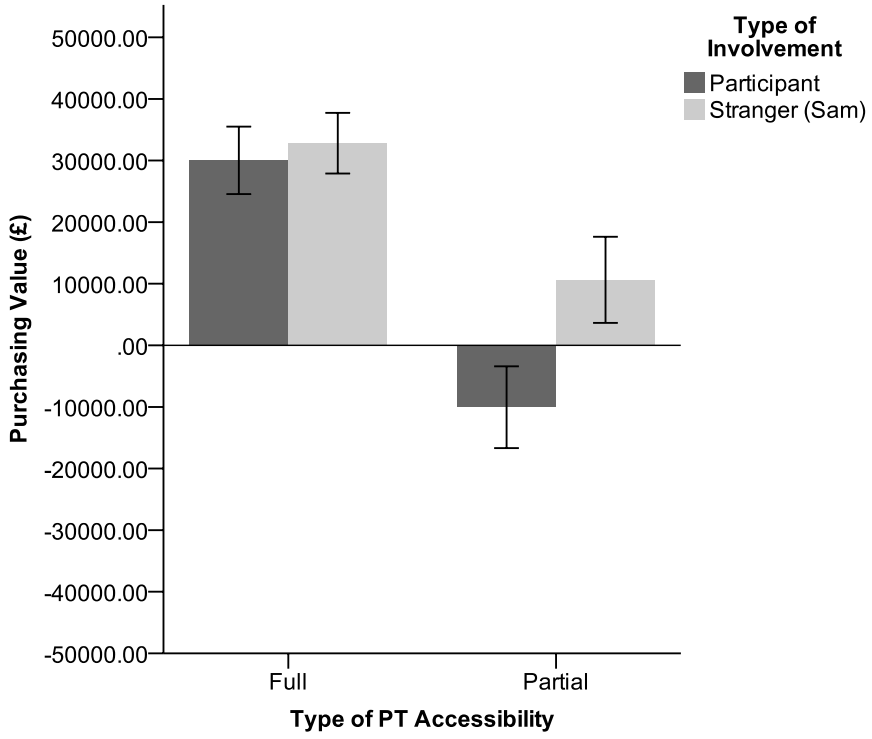
Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

2.2.2. Purchasing value

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility and type of involvement on purchasing value. The results revealed a significant main effect of both type of PT accessibility, $F(1, 296) = 104.52$, $p < .001$, $\eta_p^2 = .26$, and type of involvement, $F(1, 296) = 14.86$, $p < .001$, $\eta_p^2 = .05$, on purchasing value. Moreover, there was also a significant two-way interaction effect of type of PT accessibility by type of involvement $F(1, 296) = 8.64$, $p = .004$, $\eta_p^2 = .03$ (see Figure 3). Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

Figure 3

Participants' Reported Purchasing Values in Experiment 1a



Note. Positive purchasing values indicate utilitarian behavior (more money spent on swerve AVs than stay AVs from the budget of £50,000) and negative purchasing values indicate non-

utilitarian behavior (more money spend on stay AVs than swerve AVs from the budget of £50,000). Error bars represent 95% Confidence Intervals of the mean.

2.2.2.1. Partial PT accessibility

A follow-up simple-effect test revealed that with partial PT accessibility, the main effect of type of involvement significantly influenced purchasing value $F(1, 148) = 18.27$, $p < .001$, $\eta^2 = .11$ (see Figure 3). Specifically, participants indicated that they would pay £10,640.08 (M = £10,640; SD = £30,374.08) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved a stranger (Sam) and £10,040 (M = -£10,040; SD = £28,865.22) less for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved themselves. Accordingly, with partial PT accessibility, participants' responses were relatively more utilitarian when a stranger (Sam) was the agent in the scenario than when participant was the agent in the scenario. This difference was however eliminated when scenarios were presented with full PT accessibility (see section 2.2.2.2). This personal involvement effect is anticipated by researchers exploring binary moral decision-making in tasks providing partial information in regards to the decision-making consequences (see Greene et al., 2001; Kusev et al., 2016). For example, personal moral decisions are expected to be more emotionally salient (and cognitively demanding) than impersonal moral dilemmas, leading typically to non-utilitarian decisions. Moreover, binary moral decision-making research also demonstrated that when people are presented with the full implications of their actions, they are more likely to weigh their choices in a manner that is consistent with utilitarian ethics (Kusev et al., 2016).

2.2.2.2. Full PT accessibility

A second follow-up simple-effect test revealed that with full PT accessibility, type of involvement did not significantly influence purchasing value ($F < 1$). Accordingly, with full PT accessibility, participants indicated that they would pay £30,026.67 (M = £30,026.67; SD

= £23,804.37) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved themselves and £32,813.33 (*M* = £32,813.33; *SD* = £21,384.32) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved a stranger - Sam (see Figure 3). Therefore, with full PT accessibility (i) respondents were utilitarian in their purchasing behavior for both types of involvement (when the agent described in the scenario is a stranger - Sam or the participant in the study) and (ii) the difference in purchasing value between type of involvement was eliminated.

2.2.3. Predicting purchasing value

Two mediation analyses (by type of involvement: stranger and participant) were conducted with macro PROCESS (Hayes, 2017) to test whether the respondents' judgments of moral appropriateness mediates the relationship between type of PT accessibility and reported purchasing values. According to our mediation model, PT accessibility (as an independent variable) is the determinant of moral judgment and, in turn, this judgment (as a mediator) is the determinant of purchase decision (as the dependent variable). According to 'modern mediation analysis' (Hayes, 2017), the total effect of the independent variable on the dependent variable is analyzed for significance using a t-test and then broken down into two effects. First, the direct effect is that part of the effect of the independent variable which does not overlap with the mediator and is analyzed using a t-test. Second, the indirect effect is the multiplication of the effect of the independent variable on the mediator with the effect of the mediator (moral judgment) on the dependent variable (purchase decision). Testing with bootstrapping is used because the indirect effect is commonly not normally distributed.

The indirect effect of PT accessibility through the mediator judgments of moral appropriateness was tested with bootstrapping (*N* = 5000). We found that decision-makers' judgments of moral appropriateness was a mediator of the relationship between type of PT accessibility and purchasing values (see Table 1). Specifically, the results revealed that with

stranger involvement - Sam (Model A) and participant involvement (Model B), the standardized indirect effect of PT accessibility through the mediator judgments of moral appropriateness was significant and negative (see Table 1), indicating that participants' judgments of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and reported purchasing values. Hence, respondents' purchasing behavior was informed by their moral judgments (the utilitarian weight of moral appropriateness) and the purchasing values were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Table 1

Mediation Analysis by Type of Involvement (Experiment 1a)

				Total effect			Direct effect				95% CI(BCa)	
Model	*F(2, 147)*	*p*	R^2	*β*	*t*	*p*	*β*	*t*	*p*	Indirect effect	*LL*	*UL*
A	60.49	<.001	.45	-.39	-5.17	<.001	-.04	-.51	.608	-.35	-.482	-.243
B	86.23	<.001	.54	-.60	-9.27	<.001	-.29	-4.16	<.001	-.31	-.426	-.222

Note. Values for total effect, direct effect and indirect effect are standardized. Model A: stranger involvement (Sam). Model B: participant involvement.

The results from Experiment 1a demonstrated for the first time that with full PT accessibility (with even odds), participants were more consistent and utilitarian in their judgments of moral appropriateness and purchasing behavior than with partial PT accessibility. Moreover, one could argue that naming the stranger 'Sam' could have a confounding effect on the results (e.g., personalizing the stranger and inducing possible associations with the name Sam). Therefore, we conducted Experiment 1b with the purpose to replicate the results from Experiment 1a, but this time presenting the stranger in the scenarios anonymously (without naming the stranger).

3. Experiment 1b

3.1. Method

3.1.1. Participants

The participants in Experiment 1b (N = 300, 152 females and 148 males; mean age of the participants was 48, SD = 13.45) were recruited via an online recruitment service, and were rewarded £1 for their participation. All participants were treated in accordance with the WMA ethical code of conduct. According to power analysis, the statistical power of 2x2 ANOVA was identical to the power in Experiment 1a.

3.1.2. Experimental design, materials and procedure

Experiment 1b used the experimental design, materials and procedures from Experiment 1a, except that in the moral scenarios used for the stranger involvement condition, we did not name the stranger, as we wanted to ensure that the name does not cause potential influences on the behavioral pattern established in Experiment 1a (see Supplemental Materials for the scenarios and Figure S1 for the visual stimuli in Experiment 1b).

3.2. Results and discussion

3.2.1. Judgments of moral appropriateness

The results from a 2x2 independent measures analysis of variance revealed that type of PT accessibility significantly influenced respondents' judgments of moral appropriateness $F(1, 296) = 53.02$, $p < .001$, $\eta_p^2 = .15$. As in Experiment 1a, participants were more utilitarian in their moral judgments with full PT accessibility (M = 4.53; SD = 4.04) than with partial PT accessibility (M = 1.14; SD = 4.01). However, the results revealed that the main effect of type of involvement ($F < 1$), as well as the two-way interaction of type of involvement by type of PT accessibility ($F < 1$) on participants' judgments of moral appropriateness were not statistically significant.

3.2.2. Purchasing value

As in Experiment 1a, the results from a 2x2 independent measures analysis of variance revealed significant main effects of type of PT accessibility, $F(1, 296) = 64.94$, $p < .001$, $\eta_p^2 =$

.18, type of involvement, $F(1, 296) = 8.72$, $p = .003$, $\eta_p^2 = .03$, as well as their interaction $F(1, 296) = 4.93$, $p = .027$, $\eta_p^2 = .02$, on purchasing value. Due to the significant two-way interaction, follow-up simple-effect tests were conducted by type of PT accessibility.

3.2.2.1. Partial PT accessibility

The results from Experiment 1b confirmed the behavioral pattern identified in Experiment 1a. Specifically, participants indicated that they would pay £11,306.67 (*M* = £11,306.67; *SD* = £30,035.10) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved a stranger and £5,973.23 (*M* = -£5,973.23; *SD* = £31,672.96) less for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved themselves, $F(1, 148) = 11.75$, $p = .001$, $\eta^2 = .07$. Accordingly, as in Experiment 1a, with partial PT accessibility, participants were relatively more utilitarian when a stranger was the agent in the scenario than when participant was the agent in the scenario.

With partial PT accessibility, there was a difference in utilitarian purchasing value between participants who read scenarios about themselves and participants who read scenarios about a stranger. This difference was however eliminated when scenarios were presented with full PT accessibility (see section 3.2.2.2).

3.2.2.2. Full PT accessibility

In accordance to the findings of Experiment 1a, the results from Experiment 1b confirmed that with full PT accessibility, type of involvement did not significantly influence purchasing value ($F < 1$). With full PT accessibility, participants indicated that they would pay £28,360 (*M* = £28,360; *SD* = £26,783.23) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved themselves and £30,813.33 (*M* = £30,813.33; *SD* = £26,925.17) more for a utilitarian swerve AV than a non-utilitarian stay AV when the scenario they read involved a stranger. Accordingly, this result replicated the findings from Experiment 1a.

3.2.3. Predicting purchasing value

As in Experiment 1a, two mediation analyses (by type of involvement: stranger and participant) were conducted. According to our mediation model, PT accessibility (as an independent variable) is the determinant of moral judgment and, in turn, this judgment (as a mediator) is the determinant of purchase decision (as the dependent variable).

The indirect effect of PT accessibility through the mediator judgments of moral appropriateness was tested with bootstrapping ($N = 5000$). We found that decision-makers' judgments of moral appropriateness was a mediator of the relationship between type of PT accessibility and purchasing values (see Table 2). Specifically, the results revealed that with stranger involvement (Model A) and participant involvement (Model B), the standardized indirect effect of PT accessibility through the mediator judgments of moral appropriateness was significant and negative (see Table 2), indicating that participants' judgments of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and reported purchasing values. As in Experiment 1a, the results from Experiment 1b indicated that respondents' purchasing behavior was informed by their moral judgments (the utilitarian weight of moral appropriateness), and the purchasing values were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Table 2

Mediation Analysis by Type of Involvement (Experiment 1b)

				Total effect			Direct effect				95% CI(BCa)	
Model	*F(2, 147)*	*p*	R^2	*β*	*t*	*p*	*β*	*t*	*p*	Indirect effect	*LL*	*UL*
A	39.35	<.001	.35	-.33	-4.19	<.001	-.12	-1.64	.103	-.21	-.295	-.138
B	54.88	<.001	.43	-.51	-7.17	<.001	-.33	-4.89	<.001	-.18	-.253	-.115

Note. Values for total effect, direct effect and indirect effect are standardized. Model A: stranger involvement. Model B: participant involvement.

The results form Experiment 1b confirmed the findings of Experiment 1a; with full PT accessibility, participants were more consistent and utilitarian in their judgments of moral appropriateness and purchasing behavior than with partial PT accessibility.

However, although purchasing value indicates the monetary value participants place on each AV model, it does not capture whether the participant actually wants to buy (or indeed ride inside) an AV in the first place. We therefore conducted a second experiment to investigate whether full PT accessibility would also influence participants' willingness to buy and ride inside each AV model. Previous experimental evidence demonstrates that participants are least willing to buy a utilitarian car when they read a scenario that involved a family member inside an AV (Bonnefon et al., 2016). Therefore, in Experiment 2, it was also necessary to investigate whether this non-utilitarian buying intention remains when participants respond to full PT accessibility scenarios that involve a family member.

4. Experiment 2

4.1. Method

4.1.1. Participants

For Experiment 2, participants (N = 300) were recruited through an online recruitment service and were rewarded £1 for their participation. The sample consisted of 160 females and 140 males and the mean age was 51 (SD = 14.35). All participants were treated in accordance with the WMA ethical code of conduct. According to power analysis, the statistical power of 2x2 ANOVA was identical to the power in Experiments 1a and 1b.

4.1.2. Experimental design, materials and procedure

A 2x2 independent measures design was employed to measure the influence of type of involvement (participant or participant with family member) and PT accessibility (full or partial) on *judgments of moral appropriateness, willingness to buy* and *willingness to ride* for each AV option. All dependent variables were measured on a 10-point rating scale where 0

indicated the lowest moral appropriateness/lowest willingness. As in Experiments 1a and 1b, participants' judgments of moral appropriateness were computed as a utilitarian weight – the difference between participants' judgments of moral appropriateness for utilitarian swerve and non-utilitarian stay AVs. Similarly, the outcome variables willingness to buy and willingness to ride were computed as utilitarian weights too. Specifically, we subtracted the judgments (willingness to buy and ride) for non-utilitarian stay AVs from judgments (willingness to buy and ride) of utilitarian swerve AVs in order to generate a utilitarian weight. Therefore, positive and high difference (utilitarian weight) indicates utilitarian judgments. The first independent variable, type of involvement, was similar to the first IV in Experiments 1a and 1b, except we replaced stranger involvement with participant-and-family involvement (where both the participant and their family member are described as being agents in the scenario). The second independent variables type of PT accessibility was identical to that of Experiments 1a and 1b.

As in Experiments 1a and 1b, participants were presented with a scenario depending on the condition they were allocated to. In scenarios that involved a participant and family member, the number of pedestrians were doubled from 10 to 20. We did this, following the same logic as used in Bonnefon et al. (2016) by presenting 10 pedestrians for every 1 passenger. After reading the scenario, participants were required to make the following judgments (presented one at a time in a randomized order):

Judge the moral appropriateness of programming a car to swerve;

Judge the moral appropriateness of programming a car to stay; and

How would you rate your willingness to BUY an autonomous self-driving car programmed to swerve?;

How would you rate your willingness to BUY an autonomous self-driving car programmed to stay?;

How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to swerve?;

How would you rate your willingness to RIDE inside an autonomous self-driving car programmed to stay?

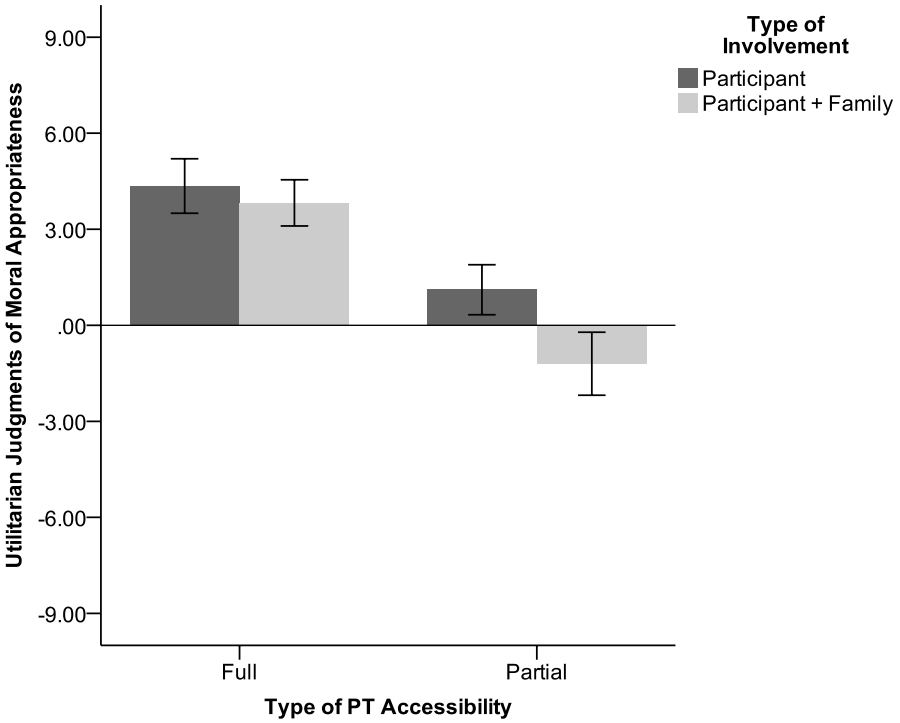
4.2. Results and discussion

4.2.1. Judgments of moral appropriateness

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on participants' judgments of moral appropriateness. The results revealed a significant main effect of type of PT accessibility, $F(1, 296) = 96.16$, $p < .001$, $\eta_p^2 = .25$, and type of involvement $F(1, 296) = 11.35$, $p = .001$, $\eta_p^2 = .04$ on participants' judgments of moral appropriateness. Moreover, the results also revealed a significant two-way interaction effect of type of involvement by type of PT accessibility on participants' judgments of moral appropriateness $F(1, 296) = 4.48$, $p = .035$, $\eta_p^2 = .02$ (see Figure 4). Accordingly, due to the significant two-way interaction, two follow-up analyses of variance were conducted by type of PT accessibility (partial and full PT accessibility).

Figure 4

Participants' Utilitarian Judgments of Moral Appropriateness in Experiment 2



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

4.2.1.1. Partial PT accessibility

A follow-up simple-effect test revealed that with partial PT accessibility, the effect of type of involvement on judgments of moral appropriateness was significant $F(1, 148) = 13.45$, $p < .001$, $\eta_p^2 = .08$. Specifically, with participant involvement, respondents' judgments of moral appropriateness were utilitarian ($M = 1.11$ $SD = 3.39$) and significantly different from the non-utilitarian judgments of moral appropriateness with participant-and-family involvement ($M = -1.20$; $SD = 4.28$); see Figure 4.

4.2.1.2. Full PT accessibility.

In contrast to the pattern of participants' judgments of moral appropriateness with partial PT accessibility, with full PT accessibility the effect of type of involvement on participants' judgments of moral appropriateness was not statistically significant ($F < 1$); see Figure 4. Accordingly, participants' judgments of moral appropriateness with participant-and-family

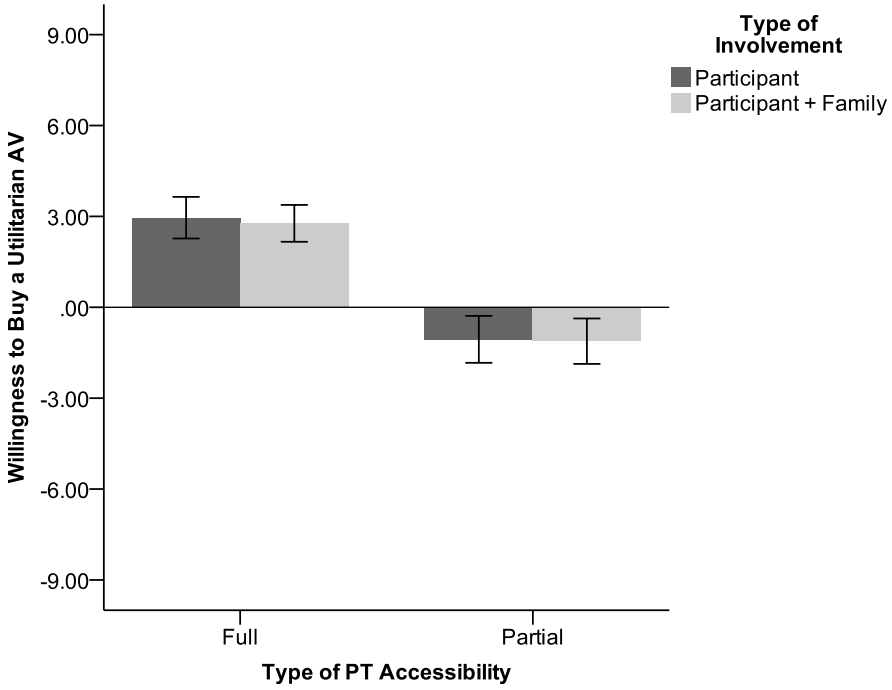
involvement (M = 3.83 SD = 3.13) as well as participants' judgments of moral appropriateness with participant involvement (M = 4.35 SD = 3.70) were both utilitarian and not statistically different.

4.2.2. Willingness to buy

A 2x2 independent measures analysis of variance was conducted to explore the influence of the independent variables (type of PT accessibility and type of involvement) on willingness to buy an AV. The results revealed a significant main effect of type of PT accessibility $F(1, 296) = 123.87$, $p < .001$, $\eta_p^2 = .30$. Specifically, with full PT accessibility, respondents' judgments of willingness to buy an AV were utilitarian (M = 2.86; SD = 2.81) and significantly different from the non-utilitarian judgments of willingness to buy an AV with partial PT accessibility (M = -1.09; SD = 3.30); see Figure 5. Moreover, the results also revealed that the main effect type of involvement ($F < 1$), as well as the two-way interaction effect of type of involvement by type of PT accessibility ($F < 1$), were not statistically significant (see Figure 5).

Figure 5

Participants' Willingness to Buy a Utilitarian AV in Experiment 2



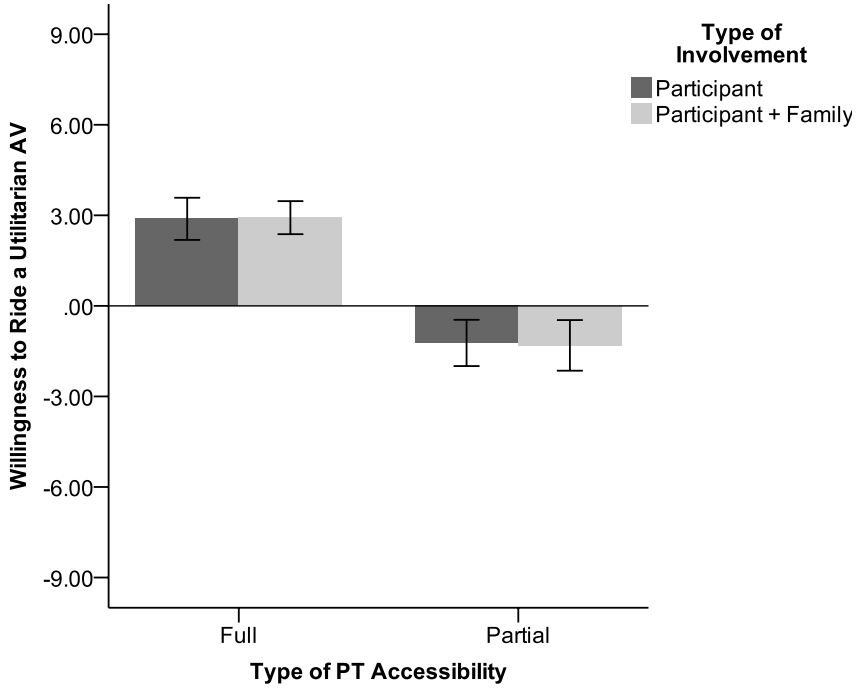
Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

4.2.3. Willingness to ride

A 2x2 independent measures analysis of variance was conducted to explore the influence of type of PT accessibility (full and partial) and type of involvement (participant and participant-and-family) on willingness to ride an AV. The results revealed that type of PT accessibility significantly influenced respondents judgments of willingness to ride an AV $F(1, 296) = 132.80$, $p < .001$, $\eta_p^2 = .31$ (see Figure 6). Accordingly, with full PT accessibility, respondents' judgments of willingness to ride an AV were utilitarian ($M = 2.90$; $SD = 2.72$) and significantly different from the non-utilitarian judgments of willingness to ride an AV with partial PT accessibility ($M = -1.27$; $SD = 3.48$); see Figure 6. Furthermore, the main effect of type of involvement ($F < 1$), as well as the two-way interaction effect of type of involvement by type of PT accessibility ($F < 1$) were statistically not significant (see Figure 6).

Figure 6

Participants' Willingness to Ride a Utilitarian AV in Experiment 2



Note. Positive mean values indicate participants' preference for utilitarian swerve AVs. Negative mean values indicate participants' preference for non-utilitarian stay AVs. Error bars represent 95% confidence intervals of the mean.

4.2.4. Predicting willingness to buy and ride

Four mediation analyses (by type of involvement and type of willingness [to buy or ride]) were conducted with macro PROCESS (Hayes, 2017) to test whether the respondents' judgments of moral appropriateness mediates the relationship between type of PT accessibility (full and partial) and reported willingness to buy and willingness to ride AVs. According to our mediation model, PT accessibility (as an independent variable) is the determinant of moral judgment and, in turn, this judgment (as a mediator) is the determinant of willingness to ride/willingness to buy (as the dependent variable).

The indirect effect of PT accessibility through the mediator judgments of moral appropriateness was tested with bootstrapping ($N = 5000$). We found that decision-makers'

judgments of moral appropriateness is a mediator of the relationship between type of PT accessibility (full and partial) and judgments for willingness to buy (Model A: participant involvement and Model B: participant with family member involvement) and willingness to ride (Model C: participant involvement and Model D: participant with family member involvement), see Table 3. The standardized indirect effect of PT accessibility through the mediator judgments of moral appropriateness was significant and negative (see Table 3), indicating that participants' judgments of moral appropriateness was a mediator of the relationship between type of PT accessibility (full and partial) and judgments for willingness to buy and willingness to ride. Hence, respondents' willingness to buy and ride judgments were informed by their moral judgments of appropriateness and were higher in the full PT accessibility condition than in the partial PT accessibility condition.

Table 3

Mediation Analysis by Type of Involvement and Type of Willingness to Buy and Ride (Experiment 2)

				Total effect			Direct effect			Indirect	95% CI (BCa)	
Model	*F(2, 147)*	*p*	R^2	*β*	*t*	*p*	*β*	*t*	*p*	effect	*LL*	*UL*
A	54.44	<.001	.43	-.54	-7.74	<.001	-.37	-5.32	<.001	-.17	-.250	-.107
B	51.40	<.001	.41	-.55	-8.02	<.001	-.33	-4.30	<.001	-.22	-.331	-.133
C	59.94	<.001	.45	-.55	-7.91	<.001	-.37	-5.42	<.001	-.18	-.254	-.119
D	59.33	<.001	.45	-.57	-8.40	<.001	-.33	-4.47	<.001	-.24	-.345	-.142

Note. Values for total effect, direct effect and indirect effect are standardized. Model A: participant involvement (willingness to buy). Model B: participant with family member involvement (willingness to buy). Model C: participant involvement (willingness to ride). Model D: participant with family member involvement (willingness to ride).

5. General Discussion

Although utilitarian AVs are expected to minimize road casualties, the way in which utilitarian AVs are described could negatively influence the public perception and adoption of

them. For instance, when promoting autonomous vehicles to potential consumers, it is particularly important to emphasize the overall benefits of utilitarian vehicles rather than the small risk of owning one (Shariff et al., 2017). Our experimental findings have demonstrated that accessibility to perspective-taking (even odds of being a passenger or pedestrian) allows people to a substantially larger extent appreciate and understand the prosocial benefits of utilitarian AVs. Accordingly, our findings from Experiments 1a, 1b and 2 revealed for the first time, that allowing human decision-makers to appreciate both situational perspectives in moral crash scenarios greatly increased their utilitarian judgments of morality, willingness to buy and ride utilitarian vehicles as well as the amount of money they spent on a utilitarian AV. In particular, our full PT accessibility task (with even odds) eliminated the behavioral inconsistency between participants' non-utilitarian purchase behavior and their utilitarian judgments of moral appropriateness (Experiments 1a and 1b). Moreover, full PT accessibility induced respondents' utilitarian prosocial purchasing behavior, utilitarian judgments of moral appropriateness and consistent utilitarian preferences across judgment tasks (Experiments, 1a, 1b and 2). Crucially, with full PT accessibility, participants' utilitarian purchasing behavior as well as their willingness to buy and ride utilitarian AVs were informed by their utilitarian moral judgments.

Moreover, in Experiment 1a, when participants read scenarios about themselves containing partial PT accessibility, they were willing to spend £10,000 more (and £6,000 more in Experiment 1b) on passenger-protective AVs than utilitarian AVs. Similar results are reported by Bonnefon et al. (2016), based on their experimental methods with limited PT. However, in Experiment 1a, this non-utilitarian pattern of purchasing value was reversed when participants read scenarios about themselves with full PT accessibility; they were willing to spend £30,000 more (and £28,000 more in Experiment 1b) on utilitarian AVs than passenger-protective models. Furthermore, with partial PT accessibility, there was a difference in utilitarian

purchasing value between participants who read scenarios about themselves and participants who read scenarios about a stranger. This difference was however eliminated when scenarios were presented with full PT accessibility (Experiments 1 and 1b). Crucially, our novel findings provide evidence that the difference in purchasing behavior between full and partial PT accessibility is informed by the difference in utilitarian moral judgments of appropriateness.

As in Experiments 1a and 1b, in Experiment 2 we found further evidence that full PT accessibility promotes and enhances utilitarian prosocial preferences and consistent behavior across judgment tasks. Specifically, with full PT accessibility scenarios, participants were more willing to buy and ride utilitarian AVs and with partial PT accessibility scenarios they were more willing to buy and ride passenger-protective AVs. This pattern of results was not influenced by the presence of a family member in the scenario (as reported by Bonnefon and colleagues, 2016). Moreover, we found evidence that the difference in respondents' judgments of willingness to buy and ride utilitarian AVs between full and partial PT accessibility, is informed by the difference in utilitarian moral judgments of appropriateness.

The link between PT and moral behavior has been established by research in both social cognition and developmental psychology (e.g., Batson, et al., 1997; Batson et al., 2003; Galinsky, et al., 2008; Galinky & Moskowitz, 2000; Kohlberg, 1976; Walker, 1980). However, our theoretical proposal and empirical results further inform PT theorists with three important findings. First, PT in morally sensitive scenarios is not always intuitive; instead, human decision-makers require accessibility to PT in order to engage in it. Second, providing decision-makers with full PT accessibility results in prosocial utilitarian decisions (such as utilitarian decisions that benefit the greatest number of people in society) that are informed by moral evaluations/judgments. Third and finally, providing decision-makers with full PT accessibility leads to greater internal preference consistency between the utilitarian AV they

judge as morally acceptable and the utilitarian AV they want for themselves (e.g., through purchasing behavior, willingness to buy and ride).

In our proposal, we provide the participants with even odds of being pedestrian or passenger in crash scenarios (eliminating the opportunity of making self-interest decisions – e.g., by selecting the 'risk averse' option, which happened to be utilitarian). This principle of impartiality offers full and accessible information regarding both situational perspectives (AV buyers can be passengers or pedestrians). Moreover, we do not induce self-interest by requiring the participants to answer the questions from a purely self-interested perspective and the tasks are not only personal as we included an independent variable type of involvement (a stranger or participant involvement).

It could be argued that full PT accessibility is a new type of VOI, which is not based on purposely induced self-interest and uneven risk options (as in Huang et al., 2019), but rather is based on even odds of being a passenger or pedestrian, and therefore with even 50/50 chance to die/live as passenger or pedestrian. Accordingly, we found that the even odds do not trigger self-interest, risk related preferences or decision biases, but facilitate opportunities for prosocial/utilitarian moral judgments and purchasing behavior. Thus, when decision-makers use full PT with even odds, they demonstrate an overall preference (internal consistency) for utilitarian AVs.

Contrary to claims that implementing a utilitarian AV policy may delay public adoption of AVs (e.g., Awad et al., 2018; Bonnefon et al., 2016; Fleetwood, 2017; Greene, 2016; Shariff et al., 2017), we argue that this does not have to be the case. Accordingly, we recommend that when car manufacturers advertise AV prosocial safety features, they should make PT in safety information fully accessible and hence increase the utilitarian AVs perceived value.

6. Conclusion

The motivation for this project came in 2016. In this collaboration we aimed to further extend our ongoing research on human moral decision-making (Kusev et al., 2016; Martin et al., 2017). In this article we argue that by default all AV buyers are pedestrians too (e.g., when crossing a road, entering/exiting their own AV), and that limiting the perspective-taking to only half of the available possibilities in crash scenarios biases participants toward expressing non-utilitarian passenger-protective purchasing preferences. Accordingly, we established that providing participants with full and accessible information regarding both situational perspectives (AVs buyers can be passengers or pedestrians) in crash scenarios boosted respondents' utilitarian prosocial purchasing behavior and consistent utilitarian preferences across judgment tasks. Crucially, with full PT accessibility, participants' utilitarian purchasing behavior as well as their willingness to buy and ride utilitarian AVs were informed by their utilitarian moral judgments. Our novel findings have, for the first time, demonstrated the impactful influence of perspective-taking accessibility on utilitarian moral decision-making.

Acknowledgement

Rose Martin: Conceptualization, Methodology, Resources, Investigation, Writing - Original Draft, Visualization, Formal Analysis. **Petko Kusev:** Methodology, Writing – Review & Editing, Formal Analysis. **Paul van Schaik:** Methodology, Writing – Review & Editing.

Declaration of Interests: None.

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*, 59-64.

Batson, C. D., Early, S., & Salvarani, G. (1997). Perspective-taking: Imagining how another feels versus imagining how you would feel. *Personality and Social Psychology Bulletin, 23*(7), 751-758.

Batson, C. D., Lishner, D. A., Carpenter, A., Dulin, L., Harjusola-Webb, S., Stocks, E. L., Gale, S., Hassan, O., & Sampat, B. (2003). "…As you would have them do unto you": Does imagining yourself in the other's place stimulate moral action? *Personality and Social Psychology Bulletin, 29*(9), 1190-1201.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21-34.

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

Cohen, J. (1988). Statistical Power for the Behavioural Sciences. Hilsdale. *NY: Lawrence Erlbaum*.

de Melo, C. M., Marsella, S., & Gratch, J. (2019). Human Cooperation When Acting Through Autonomous Machines. *Proceedings of the National Academy of Sciences*, *116*(9), 3482-3487.

Fagnant, D. J., & Kockelman, K. (2015). Preparing a nation for autonomous vehicles: barriers and policy recommendations. *Transportation Research Part A: Policy and Practice, 77*, 167-181.

Faulhaber, A. K., Dittmer, A., Blind, F., Wächter, M. A., Timm, A., Sütfeld, L. R., Stephan., A., Pipa, G., & König, P. (2018). Human decisions in moral dilemmas are largely described by utilitarianism: virtual car driving study provides guidelines for autonomous driving vehicles. *Science and Engineering Ethics*. doi: 10.1007/s11948-018-0020-x

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

Fleetwood, J. (2017). Public health, ethics and autonomous vehicles. *American Journal of Public Health, 107*(4), 532-537.

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science, 19*(4), 378-384.

Galinsky, A. D., & Moskowitz, G. B. (2000). Perspective-taking: Decreasing stereotype expression, stereotype accessibility, and in-group favoritism. *Journal of Personality and Social Psychology, 78*, 708-724.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor in mandatory ethics. *Science and Engineering Ethics, 23*(3), 681-700.

Goodall, N. J. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record: Journal of the Transportation Research Board, 2424*, 58-65.

Greene, J. D. (2016). Our driverless dilemma. *Science*, *352*(6293), 1514-1515.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108.

Hare, C. (2016). Should we wish well to all? *Philosophical Review, 125*, 451-472.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy 63*, 306-231.

Harsanyi, J. C. (1975). Can maximin principle serve as a basis for morality? A critique of John Rawl's Theory. *American Political Science Review, 69*, 594-606.

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Publications.

Huang, K., Greene, J. D., & Bazerman, M. (2019). *Proceedings of the National Academy of Science, 116*(48), 23989-23995.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697-720.

Kohlberg, L. (1976). Moral stages and moralization. In T. Lickona (Ed.), *Moral development and behavior* (pp. 31-53). London: Holt, Rinehart & Wilson.

Kollock, P. (1998). Social dilemmas: The anatomy of cooperation. *Annual Review Sociology, 24*, 183-214.

Kusev, P., van Schaik, P., Alzahrani, S., Lonigro, S., & Purser, H. (2016). Judging the morality of utilitarian actions: How poor utilitarian accessibility makes judges irrational. *Psychonomic Bulletin & Review, 23(6*), 1961-1967.

Kusev, P., van Schaik, P., Martin, R., Hall, L., & Johansson, P. (2020). Preference reversals during risk elicitation. *Journal of Experimental Psychology: General, 149,* 585- 589. http://dx.doi.org/10.1037/xge0000655

Martin, R., Kusev, I., Cooke, A., Baranova, V., van Schaik, P., & Kusev, P. (2017). Commentary: The social dilemma of autonomous vehicles. *Frontiers in Psychology, 8*(808).

Maurer, M., Gerdes, C. J., Lenz, B., & Winner, H. (2015). *Autonomous Driving: Technical, Legal and Social Aspects* doi: 10.1007/978-3-662-48847-8

Nees, M. A. (2016). Acceptance of self-driving cars: an examination of idealized versus realistic portrayals with a self-driving car acceptance scale. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 60, No. 1, pp. 1449-1453). Sage CA: Los Angeles, CA: SAGE Publications.

Nyholm, A., & Smids, J. (2016). The ethics of accident-algorithms for self-driving cars: an applied trolley problem? *Ethical Theory and Moral Practice, 19*, 1275-1289.

Rawls, J. (2009). *A theory of justice*. Harvard University Press. Original publication 1971.

Shariff, A., Bonnefon, J-F., & Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving cars. *Nature in Human Behaviour, 1*(10), 694-696.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*, 1395-1415.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*, 207-232.

Walker, L. J. (1980). Cognitive and perspective-taking prerequisites for moral development. *Child Development, 51*(1), 131-139.