Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas

Holly Wilson[1*], **Andreas Theodorou**[2]
[1]University of Bath
[2]Umeå University
hlw69@bath.ac.uk, andreas.theodorou@umu.se

Abstract

Artificially intelligent agents are increasingly used for morally-salient decisions of high societal impact. Yet, the decision-making algorithms of such agents are rarely transparent. Further, our perception of, and response to, morally-salient decisions may depend on agent type; artificial or natural (human). We developed a Virtual Reality (VR) simulation involving an autonomous vehicle to investigate our perceptions of a morally-salient decision; first moderated by agent type, and second, by an implementation of transparency. Participants in our user study took the role of a passenger in an autonomous vehicle (AV) which makes a moral choice: crash into one of two human-looking Non-Playable Characters (NPC). Experimental subjects were exposed to one of three conditions: (1) participants were led to believe that the car was controlled by a human, (2) the artificial nature of AV was made explicitly clear in the pre-study briefing, but its decision-making system was kept opaque, and (3) a transparent AV that reported back the characteristics of the NPCs that influenced its decision-making process. In this paper, we discuss our results, including the distress expressed by our participants at exposing them to a system that makes decisions based on socio-demographic attributes, and their implications.

1 Introduction

Widespread use of fully autonomous vehicles (AVs) is predicted to reduce accidents, congestion and stress [Fleetwood, 2017; Litman, 2017]. Indeed, the Institute of Electric and Electronic Engineers (IEEE) predict 75% of cars on the road will be self-driving by 2040. AVs are one of the technologies in the transportation domain most followed by the public [Beiker, 2012]. Critical to this current work; the spotlight on AVs also illuminates the 'trolley dilemma'; what action should an AV take when faced with two morally salient options? E.g. should the car hit the elderly person in the right lane, or the young child in the left?

Many argue that such scenarios are unrealistic and improbable [Brett, 2015; Goodall, 2016]. Yet, despite their improbability, such questions generate serious discussion amongst stakeholders. Germany's Ethics Commission concluded "in the event of unavoidable accident situations, any distinction based on personal features (age, gender, physical or mental constitution) is strictly prohibited" [Commission, 2017], whereas other countries are undecided on their stance. Public opinion surveys consistently report concerns about misuse, legal implications, and privacy issues [Kyriakidis *et al.*, 2015].

As one of the few morally-salient Artificial Intelligence (AI) dilemmas which has grabbed the attention of many stakeholders; policy makers, media and public, we feel this paradigm is uniquely valuable for exploring several critical research questions in human-computer interactions. These include: 1) how do our perceptions of a decision-making agent and their decision, differ dependent on whether the agent is another human or artificially intelligent; 2) how does an implementation of transparency regarding the agent's 'moral code' impact perceptions, with the expectation of calibration; 3) how does the methodology used to present such 'moral dilemma' scenarios to the public, impact their preferences and perceptions. We now outline each in turn with consideration of the current status of research and how the question can be framed within the AV scenario for further investigation.

2 Background

There are many circumstances in which decision-making Intelligent Agents (IAs) are replacing human decision-makers. Yet we have not sufficiently established how this shift in agent-type impacts our perceptions of the decision and decision-maker. The research gap is especially large in the context of morally salient decision-making. There are indications that we both inaccurately assimilate our mental model of humans with IAs, and have separate expectations and perceptions of IAs which often lack accuracy [Turkle, 2017]. We discuss the current state of research in our perceived perceptions of IA objectivity and competence, perceived moral agency and responsibility and moral frameworks. We then discuss transparency as a mechanism to calibrate inaccurate mental model of IAs. We consider crowd-sourcing moral preferences as a method to guide the moral frameworks in

*Contact Author

IAs, and how these are modulated by the methodologies used to do so.

2.1 Perceived Objectivity and Competence

Research suggests people perceive IAs as more objective and less prone to have biases than human decision makers. For example, people were found to be more likely to make decisions inconsistent with objective data when they believed the decision was recommended by a computer system than by a person [Skitka *et al.*, 1999]. Similarly, in a legal setting, people preferred to adhere to a machine advisor's decision even when the human advisor's judgment had higher accuracy [Krueger, 2016]. In the context of an AV, higher attributions of objectivity and competence could result in end-users feeling more content with decisions than they would be had the decision been made by a human driver.

2.2 Perceived Moral Agency and Responsibility

We make moral judgements and apply moral norms differently to artificial than human agents. For example, in a mining dilemma modelled after the trolley dilemma, robots were blamed more than humans when the utilitarian action was not taken [Malle *et al.*, 2015]. This action was also found to be more permissible with the robot than the human; robots were expected to make utilitarian choices. This could have implications for the moral frameworks we might program into machines—which might not necessarily be equivalent to the frameworks we prescribe to humans. The impact agent type has on responsibility attribution is similar. After reading an AV narrative, participants assigned less responsibility to an AV at fault than to a human driver at fault [Li *et al.*, 2016]. We initially have higher expectations of IAs, yet are less forgiving when things go wrong, attributing more blame.

2.3 Inaccurate Mental Models of Moral Framework

When we form mental models about a newly encountered IA, we draw on past experiences, clues from an object's physical characteristics and may anthropmorphise. Therefore, previously encountering an IA which is physically similar but operates on deontological principles can be misleading. Based on this past experience, we may assume this newly encountered IA also operates on deontological principles. Alternatively, due to anthropomorphism of the IA, we may assume human bias mechanisms. If in fact, the newly encountered IA is embedded with a utilitarian moral framework, then we form an inaccurate mental model. We would then have reduced understanding of and perhaps even a sub-optimal interaction with the IA.

Alternatively, an IA may appear too dissimilar to a person for an observer to attribute it a moral framework at all. This too leads to difficulties in predicting the IA's behaviour. For optimal interaction and to make informed choices about usage, we require accurate mental models of moral frameworks. Yet, there are no previous studies which explicitly investigate the impact of implementing IA moral framework transparency on human perceptions of that IA.

2.4 Transparency to Calibrate Mental Models

Transparency is consider an essential requirement for the development of safe-to-use systems [Theodorou *et al.*, 2017]. A careful implementation of transparency can, for example, enable real-time calibration of trust to the system [Dzindolet *et al.*, 2003]. Wortham *et al.* [2017] revealed a robot's drives, competences and the actions of a robot to naive users through the usage of real-time AI visualisation software; this additional information increased accuracy of observers' mental models for the robot. Although, notably, transparency did not result in perfect understanding: some still overestimated the robot's abilities. The present research carries on our exploration of transparency for IAs.

2.5 Crowd-Sourcing Moral Preferences

AVs *could*, but not necessarily *should*, be programmed with behaviours that conform to a predetermined moral framework such as utilitarian, deontological or with a normative framework. There has already been valuable work garnering normative preferences for the AV moral dilemma; participants given narratives of different dilemmas, showed a general preference to minimise casualty numbers rather than protecting passengers at all costs [Bonnefon *et al.*, 2016] . However, people no longer wished to sacrifice the passenger when only one life could be saved, an effect which was amplified when an additional passenger was in the car such as a family member.

Awad *et al.* [2018] used a Massive Online Experiment named 'The Moral Machine' to determine a global moral preference for the AV—trolley dilemma: users selected between two options which were represented by a 2D, pictorial, birds eye view as a response to 'What should the self-driving car do?'. This work made an extensive contribution to establishing global normative preferences as well as finding cross-cultural ethical variation in preference.

An interesting extension upon 'The Moral Machine' foundation, is to explore how decision-making may differ when a dilemma is presented in a more immersive medium. When viewing pictures or reading narratives, as in the study of Awad [2017], there is less emotional elicitation than in the equivalent real life situations, whereas VR has higher ecological validity, provoking true to life emotions and behaviours [Rovira *et al.*, 2009]. Importantly, people have been found to make different decisions for moral dilemmas in immersive VR simulation than in desktop VR scenarios [Pan and Slater, 2011]. Specifically, immersive VR induced more panic, and less utilitarian decision making.

3 Technology Used

The observation that moral intuitions may vary with presentation motivates us to use a VR environment for our present work. Additionally, participants are passengers inside the car, rather than bystanders removed from the dilemma. Unlike most past research, transparency in this study will be implemented post-decision rather than real-time. We developed a VR environment in Unity, optimised for Oculus Rift. Unity software was chosen due to the wide range of free available assets. The AV simulation, a screenshot is seen in figure 1,

is designed so that participants are seated in the driver's seat of a car. The car has detailed interior to increase realism and thus immersion. The car, positioned on the left hand lane as the experiments took place in the UK, drives through a city environment and approaches a zebra-crossing. There are non-playable characters, of the various "physical types" described in Section 3, crossing the road. There are eleven scenarios in total, of which one is a practice and thus devoid of characters.

Figure 1: Participant is seated as passenger. On the crossing ahead there is a pair who differ in body size.

We developed a VR AV simulation to explore public perceptions of moral decisions[1]. This roughly simulates the Moral Machine scenarios [Awad, 2017], in which an AV hits one of two individuals or groups of pedestrians. This experimental tool facilitated two experiments presented here, which seek to answer questions posed above.

In this section, we first present our decision-making framework. A brief outline of our VR simulation is provided, alongside justification for design choices, selection of pedestrian attributes and how transparency is implemented. We then move to outline the experimental design of the two experiments.

Design of Moral Dilemma and AI Decision Making We opted to use only a selection of the dimensions used in other studies on moral preferences. This is because we are not measuring which characteristics the participants would prefer to be saved, but rather the response to the use of characteristic based decision-making in the first place. We picked the three more visible characteristics: occupation, sex, and body size; due to limited availability of assets and the pictorial rather than textual presentation of the scenario to the participant.

Occupation includes four representative conditions: a medic to represent someone who is often associated with contribution to the wealth of the community, military to represent a risk-taking profession [McCartney, 2011], businessman or businesswoman as it is associated with wealth, and finally unemployed. The body size can be either non-athletic or athletic slim. To further reduce the dimensions of the problem, we used a binary gender choice (female and male). Although we varied race *between* scenarios (Caucasian, Black, and Asian), the character pairs *within* scenarios were always of the same race. Examples of characters used are depicted in Fig. 2. Note, we do not claim that this is the 'right' hierarchy of social values —or that a choice should take place based on socio-demographic characteristics in the first place. Rather, we simply use this hierarchy as a mean to investigate people's perception of morally salient actions taken by AI systems.

[1]Code for this simulation will be made available on publication.

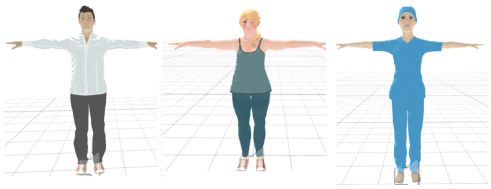

Figure 2: Examples of characters used. From left to right: Asian slim businessman, Caucasian non-athletic unemployed female, and asian female athletic-slim medic.

4 Experimental Design

We ran a study with three independent groups; human driver (Group 1), opaque AV (Group 2), and transparent AV (Group 3). We randomly allocated participants to the independent variable conditions. For each condition, both the experimental procedure and the VR Moral Dilemma Simulator were adjusted in the pre-treatment briefing. In this section, we describe our procedure for each condition.

4.1 Participants recruitment and pre-briefing

To reduce an age bias often observed in studies performed with undergraduate and postgraduate students, we decided to recruit through a non-conventional means. Participants recruitment took place at a local prominent art gallery, where we exhibited our VR simulation as part of a series of interactive installations. Ethics approval was obtained from the Department of Computer Science at University of Bath. Members of the public visiting the gallery were approached and invited to take part to the experiment. They were told the purpose of the experiment is to investigate technology and moral dilemmas in a driving paradigm. After completing a preliminary questionnaire to gather demographic, driving-preference and social-identity data, participants entered the VR environment. After either completion of the VR paradigm or when the participant decided to stop, the participant was requested to fill out a post simulation questionnaire. This questionnaire aims to capture the participant's perceptions of the agent controlling the car. It includes dimensions of likeability, intelligence, trust, prejudice, objectivity and morality. Whilst some questions are hand-crafted for the purposes of this study, most are derived from the GodSpeed Questionnaire Series as they are demonstrated to have high internal consistency [Bartneck *et al.*, 2009]. The majority of items are measured on a 5–point Likert Scale.

4.2 Human Driver Condition

Participants were informed that they were to be a passenger sat in the driver seat, in either an AV or a car controlled by a human driver. In reality, the same intelligent system controlled the car in both conditions. To make the human driver condition believable, before putting on the headset, the participants were shown a 'fake' control screen. A physical games controller, which the experimenter pretended to 'use' to control the car, was placed at the table. At the end of the experiment, participants were debriefed and told that there was no

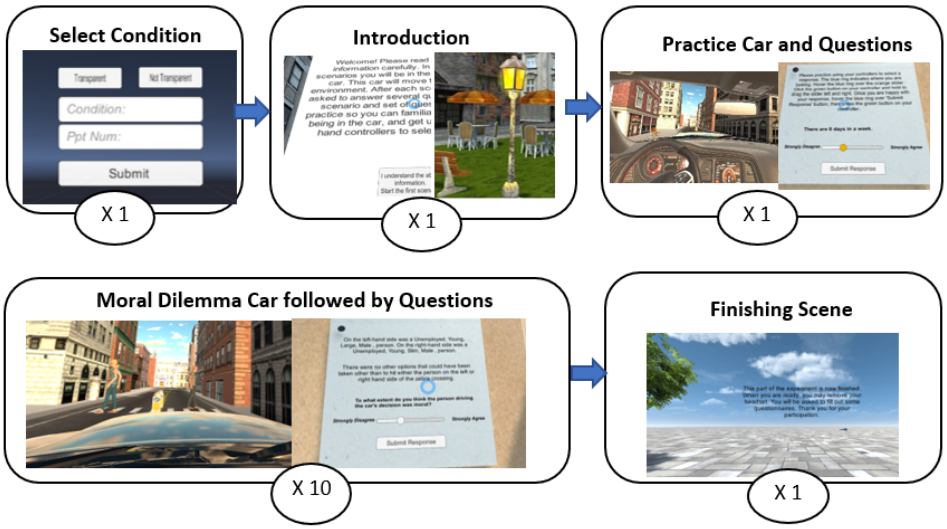

Figure 3: The Preliminary Condition Scene is followed by an Introduction Scene, the Practice Car and Practice Questions Scene. The Scenario followed by the Question Scenes are then cycled through ten times, for ten different moral dilemmas. The Finishing Scene closes.

actual human controlling the car and it was automated as in the AV condition.

4.3 Opaque and Transparent AV Conditions

Transparency here, refers to revealing to the end user the moral framework the agent uses to make its decision. The moral framework for this paradigm is social value. The difference between the transparent and non-transparent condition is in the Question Scene. Post-scenario, after the AV has hit one of two pedestrians, a statement is made that "The self-driving car made the decision on the basis that…" then the reasoning logic is inserted next. For example, if the pair consisted of one medic and another military, the justification will state "Medics are valued more highly than military, business or undefined professions". Whereas, if the pair differ only in gender, it will state "Both sides have the same profession and body size, however females are valued more highly than males". In this experiment, the transparency only relays aspects of the agent's moral framework. There is no transparency over mechanics, such as whether the brakes were working, the speed of the car, or turning direction.

Several modalities of transparency implementation were considered such as diagrams, design metaphors and audio, although written depictions were ultimately used. Post-decision transparency was chosen to be appropriate, as this paradigm invokes a fast paced situation where real-time implementation is infeasible due to technical and human processing limitations.

5 Results

Imbalance of baseline variables is usually considered undesirable, as the essence of a controlled trial is to compare groups that differ only with respect to their treatment. Others suggest that randomised—unbalanced—trials provide more meaningful results as they compact chance bias [Roberts and Torgerson, 1999]. A Chi-squared test of goodness-of-fit was performed to determine frequencies of gender, ethnicity, age, driving preferences and programming experience between the three conditions (see 1). Groups were found to be unbalanced for gender and ethnicity. The ethnicity difference between the groups is due to a number of people who did not answer the 'Ethnicity' question; the vast majority of all groups consisted of participants who identified themselves as *white* and no other ethnicities were reported. The unbalance for gender, however, should be taken into consideration during the analysis of the results.

Variable	Group 1: Human Driver	Group 2: Opaque AV	Group 3: Transparent AV	*X(2)*	*P*
Gender Male	5	5	14		
Gender Female	12	11	4		
Gender Unknown	1	0	0		
				13.89	**0.03**
White	16	14	17		
Asian	0	0	0		
Black/Caribbean	0	0	0		
None/Unknown	2	2	1		
				27.66	**0.001**
16-17	1	1	0		
18-25	2	5	3		
26-35	5	3	6		
36-45	3	3	0		
46-60	6	4	5		
60+	1	0	4		
				15	0.45
Automatic	2	2	2		
Manual	5	6	10		
Both	4	3	4		
None/Unknown	7	5	2		
				10.23	.33
Program	5	6	7		
Do not program	12	10	11		
Unknown	1	0	0		
				5.03	.54
Total Participants	18	16	18		

Table 1: Participants' Demographics. Groups were found to be unbalanced for gender and ethnicity.

For comparisons of human-driver versus the opaque AV, a

one-way ANOVA was conducted on all ordinal Likert scale variables. All but two associations were found to be non-significant (n.s.), see Table 2. The AV was perceived to be significantly less human-like (p = 0.001), and less morally culpable (p = 0.04) than the human driver. Although the impact of agent-type was n.s., medium effect sizes were found for the human driver being perceived as more pleasant (η_p^2 = 0.105, d = 0.69) and nice (η_p^2 = 0.124), d = 0.75) than the AV [Becker, 2000]. In a second one-way ANOVA, comparing the opaque AV and transparent AV conditions, three significant effects were found. The AV was perceived to be significantly more unconscious rather than conscious ($p < 0.001$), machine-like than humanlike ($p = 0.04$) and intentional rather than unintentional ($p = 0.038$) (see Table 5) in the transparent condition than the non-transparent condition. All other differences were n.s. In a third one-way ANOVA, comparing the human driver and transparent AV condition, four significant effects were found see Table 3. The human driver was found to be significantly more pleasant ($p = 0.01$), nice ($p = 0.018$), humanlike ($p < 0.001$) and conscious ($p < 0.001$) than the transparent AV. All other differences were n.s.

Question	N	Mean (*SD*)	t (*df*)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Machinelike - Humanlike					
Group 1: Human Driver	17	3.2 (0.97)			
Group 2: Opaque AV	16	2.1 (0.96)			
			3.42 (31)	**0.001**	0.191
Unpleasant - Pleasant					
Group 1: Human Driver	16	3 (=0.35)			
Group 2: Opaque AV	17	2.6 (0.89)			
			1.38 (31)	0.18	0.105
Awful - Nice					
Group 1: Human Driver	17	3 (=0)			
Group 2: Opaque AV	16	2.6 (0.89)			
			1.53 (31)	0.13	0.124
Culpability and Blame					
Morally Culpable (Scale 1-4)					
Group 1: Human Driver	16	3.37 (0.7)			
Group 2: Opaque AV	16	2.56 (1.21)			
			-2.07 (30)	**0.04**	0.18
Blame (Scale 1-4)					
Group 1: Human Driver	15	2.07 (0.7)			
Group 2: Opaque AV	16	2.44 (1.21)			
			-0.94 (29)	0.354	0.020

Table 2: Perceptions based on type of agent Human Driver v Opaque AV: The results show that participants in Group 2 perceived the AV as significantly more machinelike compare to participants in Group 1. Moreover, participants in the opaque AV condition described the robot as less morally culpable compared to the ones in Group 1.

A chi-square test of independence was performed to examine the relation between transparency and understanding of the decision made $\chi^2(1) = 7.34p = 0.007$. Participants in the transparent condition were more likely to report understanding (87.5%) than (43.75%) (see 4).

The majority of participants across conditions expressed a preference for decisions made in moral dilemmas to be made at random rather than on the basis of social-value. Preferences are as follows; 71.7% random, 17.9% social value, 7.7% unspecified criteria and 2.6% preferred neither (Fig. 4).

6 Discussion

We investigated how certain factors, namely agent type and transparency level, impact perceptions of a decision maker and the decision made in moral dilemmas. We discuss the findings for these conditions and consider the qualitative and quantitative findings that emerged from both. We place initial emphasis on participants' reactions to the decision being made on social value and the modulating impact of methodology.

Question	N	Mean (*SD*)	t (*df*)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Unpleasant - Pleasant					
Group 1: Human Driver	17	3.0 (0.35)			
Group 3: Transparent AV	17	2.35 (0.93)			
			2.68 (32)	**0.01**	0.183
Awful - Nice					
Group 1: Human Driver	17	3.0 (0.0)			
Group 3: Transparent AV	17	2.47 (0.87)			
			2.5 (32)	**0.018**	0.163
Machinelike - Humanlike					
Group 1: Human Driver	17	3.24 (0.97)			
Group 3: Transparent AV	18	1.5 (0.92)			
			5.42 (33)	**0.000**	0.47
Unconscious - Conscious					
Group 1: Human Driver	17	3.0 (1.17)			
Group 3: Transparent AV	18	1.33 (0.59)			
			5.35 (33)	**0.000**	0.464

Table 3: Perceptions based on type of agent; comparing Human Driver to the the Transparent AV. Participants in the Human Driver condition described their driver as significantly more pleasant than participants of the Transparent AV condition described the AV's behaviour. In addition, participants in Group 3 perceived the Transparent AV as less nice than the subjects in Group 2. Not surprisingly, Group 1 also described the Human driver as more humanlike compare to Group 3 which described the AV as machinelike. Moreover, participants in the Human Driver condition significantly perceived their driver as conscious compare to subjects in Group 3.

Question	N	Mean (*SD*)	t (*df*)	p	η_p^2
Objectivity (Scale 1-5)					
Deterministic - Undeterministic					
Group 1: Human Driver	17	2.89 (1.11)			
Group 3: Transparent AV	17	2.0 (1.0)			
			2.43 (32)	**0.02**	0.156
Unpredictable - Predictable					
Group 1: Human Driver	17	3.06 (1.34)			
Group 3: Transparent AV	18	4.0 (1.29)			
			-2.12 (33)	**0.04**	0.120
Intentional - Unintentional					
Group 1: Human Driver	17	3.09 (1.14)			
Group 3: Transparent AV	18	1.83 (1.2)			
			3.09 (33)	**0.004**	0.224
Culpability and Blame					
Morally Culpable (1-4)					
Group 1: Human Driver	16	3.37)			
Group 3: Transparent AV	18	3.05 (1.3)			
			-3.89 (32)	**0.00**	0.321
Blame (1-5)					
Group 1: Human Driver	15	2.07 (0.7)			
Group 3: Transparent AV	18	3.0 (1.28)			
			-2.52 (31)	**0.02**	0.169

Table 4: Perceptions based on type of agent; comparing Human Driver to the the Transparent AV. Subjects in the Human Driver condition significantly described their driver as more deterministic in its decision than participants in the Transparent AV condition. Moreover, Group 3 found the Transparent AV more Predictable compared to participants in Group 1. Group 1 considered the Human Driver's actions significantly more Intentional than participants in the Transparent AV condition did. Furthermore, experimental subjects in the Human Driver condition perceived the driver as less morally culpable and assigned less blame to the driver than participants in Group 3 did to the AV.

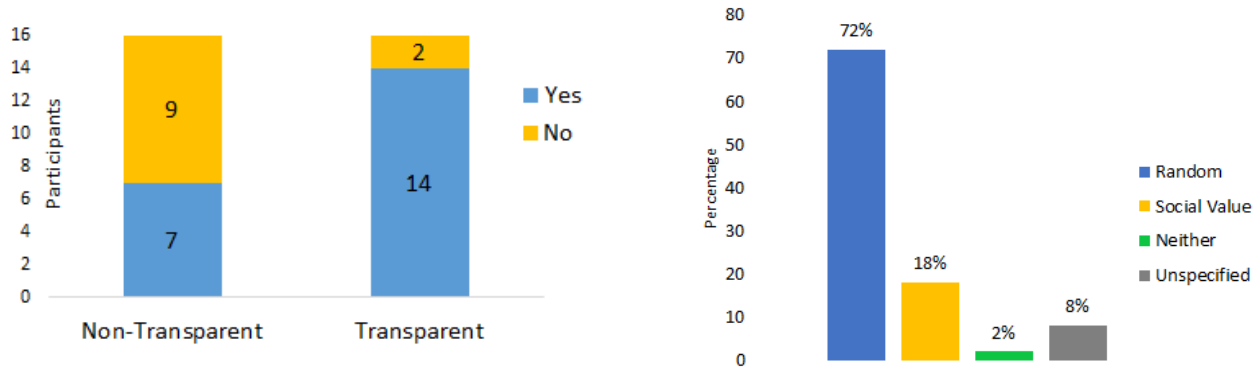

Figure 4: Left: more participants self-reported understanding the decision made by the AV in the transparent condition than in the non-transparent condition. Right: participants' preferences for the decision an agent makes when faced with hitting one of two pedestrians after a virtual reality simulation. Choices include: selecting between pedestrians at random, basing the decision social value, neither or an alternate option generated by the participant

Question	N	Mean (SD)	t (df)	p	η_p^2
Godspeed Questionnaire (Scale 1-5)					
Machinelike - Humanlike					
Group 2: Opaque AV	16	3.2 (0.97)			
Group 3: Transparent AV	18	2.1 (0.96)			
			-2.1 (32)	**0.04**	.084
Unconscious - Conscious					
Group 2: Opaque AV	16	2.75 (1.34)			
Group 3: Transparent AV	18	1.33 (0.59)			
			-4.09 (32)	**0.001**	0.294
Intentional - Unintentional					
Group 2: Opaque AV	16	2.69 (1.25)			
Group 3: Transparent AV	18	1.83 (1.2)			
			-2.13 (32) w	**0.038**	0.082

Table 5: Perceptions based on level of transparency: The autonomous car was perceived by participants in the non-transparent AV condition to be significantly more 'Humanlike' than subjects in Group 3. Moreover, participants in Group 3 found the AV to be significantly more 'Unconscious' rather than 'Conscious' compared to participants in Group 2. Finally, the Group 3 participants described the actions by the AV significantly more 'Intentional' than subjects the non-transparent condition. No other significant results were reported.

6.1 Selection Based on Social Value

Our experiment elicited strong emotional reactions in participants, who vocalised being against selection based on social value. This response was far more pronounced in the autonomous vehicle condition than with the human driver. Our quantitative and qualitative data raise interesting questions about the validity of data captured by Trolley Problem experiments, such as the the Moral Machine [Shariff *et al.*, 2017; Awad *et al.*, 2018] as a means to 'crowdsource' the moral framework of our cars by using socio-economic and demographic data. While such data are definitely worth analysing as a means to understand cultural differences between populations, they may not necessarily be representative of people's preferences in an actual accident. A lack of an option to make an explicit 'random' choice combined with the use of a non-immersive methodologies, could lead participants in 'text description' conditions to feel forced to make a logical choice.

The disparity in findings reflects differing processes of decision making between the rational decision making in the Moral Machine and emotional decision-making in the current experiment. Due to their increased realism, as previously discussed, VR environments are known to be more effective at eliciting emotion than narratives or 2D pictures. Although the graphics used in this experiment were only semi-realistic, the screams were real recordings. Participants commented on the emotional impact and stress the screams had on them. Additionally, they were visibly upset after completing the experiment and expressed discomfort at having to respond about social value decisions of which they disagreed with on principle. Other participants removed their consent, requested data to be destroyed, or even provided us with strongly-worded verbal feedback. Likely, the emotion elicitation was enhanced further, as the participant was a passenger inside the car as opposed to a bystander removed from the situation as in past experiments. It is unlikely that the Moral Machine and other online-survey narrative-based moral experiments elicit such emotional responses. This is also supported by Pedersen *et al.* [2018], where participants in autonomous-vehicle simulation study significantly altered their perception of the actions taken by an AV when a crash could lead to real-life consequences.

Our qualitative results also indicate that subjects may feel uncomfortable being associated with an autonomous vehicle that uses protected demographic and socio-economic characteristics for its decision-making process. This might be due to a belief that the users of such a product will be considered as discriminators by agreeing with a system that uses gender, occupation, age, or race to make a choice. This belief could potentially also lead to a fear that the user may share any responsibility behind the accident or be judged by others —including by the experimenter.

6.2 Perceptions of Moral Agency

Based on past research, we predicted that the autonomous car condition would be perceived as more objective and intelli-

gent but less prejudiced, conscious and human-like, and be attributed less culpability and moral agency than the 'human driver'. We found that human drivers were perceived as significantly more humanlike and conscious than autonomous cars. This finding is consistent with expectations and validates that participants perceived the two groups differently, especially, as we primed our subjects in the pre-briefing by telling them that the driver is a 'human'.

Human drivers (Group 1) were perceived to be significantly more morally culpable than autonomous driver in the opaque AV condition (Group 2). However, strikingly, the reverse was observed when the car's decision-making system was made transparent. Furthermore, in the transparency condition, participants assigned significantly more blame to the car than the 'human' driver. Although, as Group 1 believed the experimenter was controlling the car, less blame may be due to identification with the experimenter or other person specific confounds. Our implementation of transparency made the machine nature of the AV explicitly clear to its passengers, with participants in Group 3 (transparency condition) describing the AV as significantly more machinelike compared to participants in Groups 1 and 2. Our findings contradict recent work by Malle *et al.* [2016], which demonstrate people perceive mechanistic robots as having less agency and moral agency than humans. Moreover, our results conflict with the results presented in Li *et al.* [2016], where participants assigned less responsibility to an autonomous vehicle car at fault than to a human driver at fault.

In the transparency condition we made the passengers aware that the car used demographic and social-value characteristics to make a non-random decision. This explains why participants in Group 3 also significantly described the AV as more intentional rather than unintentional compared to subjects in the other two conditions. Although we inevitably unconsciously anthropomorphise machines, something that our post-incident transparency minimised by significant reducing its perception as humanlike and as conscious, we still associate emotions more easily with humans than machines [Haslam *et al.*, 2008]. Reduced emotion in decision-making is linked to making more utilitarian judgements, as supported by behavioural and neuropsychological research [Moll and de Oliveira-Souza, 2007; Lee and Gino, 2015]. Therefore, we believe that participants in the transparency condition may have also perceived decisions as utilitarian, as the car was maximising the social value —at least based on same perception— it would save.

We believe that the increased attribution of moral responsibility is due to realisation that the action was determined based on social values, something that subjects (across all groups), as we already discussed, disagreed with. This is supported by past research findings: we perceive other humans as less humanlike when they lack empathy and carry out actions which we deem to be morally wrong. For example, offenders are dehumanised based on their crimes, which we view as 'subhuman' and 'beastly' [Bastian *et al.*, 2013]. Actions that go against our moral codes can elicit visceral responses which is consistent with the emotional reactions of the participants of the current study.

Our findings may also reflect forgiveness towards the 'human' driver or even the opaque AV, but not the transparent AV. This is supported by previous studies from the literature, which demonstrate how we tend to forgive human-made errors easier than machine-made errors [Madhavan and Wiegmann, 2007; Salem *et al.*, 2015]. This effect is increased when the robot is perceived as having more autonomy [Kim and Hinds, 2006]. In addition, Malle *et al.* [2015] demonstrate, with the use of a moral dilemma modelled after the trolley problem, that robots are blamed more than humans when a utilitarian action is not taken. Furthermore, their results also suggest that a utilitarian action is also be more permissible —if not expected— when taken by a robot. If for example the robot was performing random choices, then the moral blame might have been higher.

The gender imbalance between the groups might also be a factor, but potentially not a conclusive one. The Moral Machine dataset shows minor differences in the preferences between male-identified and female-identified participants [Awad *et al.*, 2018], e.g. male respondents are 0.06% less inclined to spare females, whereas one increase in standard deviation of religiosity of the respondent is associated with 0.09% more inclination to spare humans. Further analysis by Awad [2017] indicates that female participants were acting *slightly* more utilitarian than males —but both genders are acting as such. Group 3 was the only group where the vast majority of its members identified themselves as males and some of its members may have disagreed with the actions taken by the agent. Whilst a plausible explanation, it does not discount the previous discussions —especially, considering that males in the Moral Machine still had a preference towards utilitarian actions. Still, we recognise the need to recapture the data for Group 3.

6.3 Mental Model Accuracy

Although this was not the focus of the study, we asked participants from Groups 2 and 3 (opaque and transparent AV respectively) to self-evaluate their understanding of how a decision was made. Significantly more participants in the transparency condition reported an understanding of the decision-making process. In addition, passengers in the transparent AV also rated the AV as significantly more predictable than the 'human' driver and higher (non-significant result; Mean for Group 2 is 3.31 and mean for Group 3 is 4) than the opaque AV.

Having accurate mental models by having an understanding of the decision-making mechanism is crucial for the safe use of the system. In this experiment we used a post-incident implementation of transparency instead of a real-time one. Hence the user could only calibrate its mental model regarding the decision and the agent *after* the incident. However, as the user repeated the simulation ten times, she could still use previously gathered information, e.g. that the car makes a non-random decision or even of the priorities of the AV's action-selection system, and predict if the car would change lanes or not.

7 Conclusions and Future Work

Exciting new technology is stirring up debates which speak to ethical conundrums and to our understanding of human

compared to machine minds. By focusing our efforts on understanding the dynamics of human-machine interaction, we can aspire to have appropriate legislation, regulations and designs in place before such technology hits the streets. In this project we created a moral-dilemma virtual-reality paradigm to explore questions raised by previous research. We have demonstrated morally salient differences in judgement based on very straightforward alterations of presentation. Presenting a dilemma in VR from a passenger's view gives an altered response versus previously reported accounts from a bird's eye view. In this VR context, presenting the same AI as a human gives a completely different set of judgements of decisions versus having it presented as an autonomous vehicle, despite the subjects' knowing in both cases that their environment was entirely synthetic.

There are important takeaway messages to this research. Crowd-sourced preferences in moral-dilemmas are impacted by the methodology used to present the dilemma as well as the questions asked. This indicates a need for caution when incorporating supposed normative data into moral frameworks used in technology. Furthermore, our results indicate that the show of transparency makes the agent appear to be significantly less anthropomorphic. In addition, our results agree with the literature that transparency can help naive users to calibrate their mental models. However, our results also show that transparency alone is not sufficient to ensure that we attribute blame—and, therefore, responsibility—only to legal persons, i.e. companies and humans. Therefore, it is essential to ensure that we address by ownership and/or usage our responsibility and accountability [Bryson and Theodorou, 2019].

Here, it is important to also recognise a limitation of our own study; the lack of a 'self-sacrifice' scenario, where the car sacrifices its passenger to save the pedestrians. The implementation of this 'self-sacrifice' feature could potentially lead to different results. A missed opportunity is that we did not collect users' preferences at each dilemma to enable further comparisons. Finally, a future rerun to both gather additional data and eliminate any concerns for results due to gender imbalance between the groups is necessary.

Acknowledgments

We would like to acknowledge Joanna Bryson for her guidance with the experimental design and feedback on this paper, Leon Watts for lending us the necessary computer equipment to conduct our study, Alan Hayes for his feedback, and The Edge Arts Centre for hosting us during data collection. Thanks also to the helpful reviewers. We also acknowledge EPSRC grant [EP/S515279/1] for funding Wilson. Theodorou was funded by the EPSRC grant [EP/L016540/1]. Final publication and dissemination of this paper would not have been possible with out the European Union's Horizon 2020 research and innovation programme under grant agreement No 825619 funding Theodorou.

References

[Awad *et al.*, 2018] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59, 2018.

[Awad, 2017] Edmond Awad. *Moral machines: perception of moral judgment made by machines*. PhD thesis, Massachusetts Institute of Technology, 2017.

[Bartneck *et al.*, 2009] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009.

[Bastian *et al.*, 2013] Brock Bastian, Thomas F Denson, and Nick Haslam. The roles of dehumanization and moral outrage in retributive justice. *PloS ONE*, 8(4):e61842, 2013.

[Becker, 2000] Lee A Becker. Effect size (es). *Retrieved September*, 9:2007, 2000.

[Beiker, 2012] Sven A Beiker. Legal aspects of autonomous driving. *Santa Clara L. Rev.*, 52:1145, 2012.

[Bonnefon *et al.*, 2016] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576, 2016.

[Brett, 2015] Rose Brett. The myth of autonomous vehicles' new craze: Ethical algorithms, November 2015.

[Bryson and Theodorou, 2019] Joanna Bryson and Andreas Theodorou. How Society Can Maintain Human-Centric Artificial Intelligence. In Marja Toivonen-Noro, Evelina Saari, Helinä Melkas, and Mervin Hasu, editors, *Human-centered digitalization and services*. 2019.

[Commission, 2017] Ethics Commission. Automated and connected driving, 2017.

[Dzindolet *et al.*, 2003] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human Computer Studies*, 58(6):697–718, 2003.

[Fleetwood, 2017] Janet Fleetwood. Public health, ethics, and autonomous vehicles. *American Journal of Public Health*, 107(4):532–537, 2017.

[Goodall, 2016] Noah J Goodall. Away from trolley problems and toward risk management. *Applied Artificial Intelligence*, 30(8):810–821, 2016.

[Haslam *et al.*, 2008] Nick Haslam, Yoshihisa Kashima, Stephen Loughnan, Junqi Shi, and Caterina Suitner. Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social Cognition*, 26(2):248–258, 2008.

[Kim and Hinds, 2006] Taemie Kim and Pamela Hinds. Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pages 80–85, 2006.

[Krueger, 2016] Frank Krueger. Neural signatures of trust during human-automation interactions. Technical report, George Mason University Fairfax United States, 2016.

[Kyriakidis *et al.*, 2015] Miltos Kyriakidis, Riender Happee, and Joost CF de Winter. Public opinion on automated driving: Results of an international questionnaire among 5000 respondents. *Transportation Research Part F: Traffic Psychology and Behaviour*, 32:127–140, 2015.

[Lee and Gino, 2015] Jooa Julia Lee and Francesca Gino. Poker-faced morality: Concealing emotions leads to utilitarian decision making. *Organizational Behavior and Human Decision Processes*, 126:49–64, 2015.

[Li *et al.*, 2016] Jamy Li, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram F Malle. From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report, SAE Technical Paper, 2016.

[Litman, 2017] Todd Litman. *Autonomous vehicle implementation predictions*. Victoria Transport Policy Institute, 2017.

[Madhavan and Wiegmann, 2007] Poornima Madhavan and Douglas A Wiegmann. Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4):277–301, 2007.

[Malle *et al.*, 2015] Bertram F Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 117–124. ACM, 2015.

[Malle *et al.*, 2016] Bertram F Malle, Matthias Scheutz, Jodi Forlizzi, and John Voiklis. Which robot am i thinking about?: The impact of action and appearance on people's evaluations of a moral robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 125–132. IEEE Press, 2016.

[McCartney, 2011] Helen McCartney. Hero, Victimor Villain? The Public Image of the British Soldier and its Implications for Defense Policy. *Defense & Security Analysis*, 27(1):43–54, mar 2011.

[Moll and de Oliveira-Souza, 2007] Jorge Moll and Ricardo de Oliveira-Souza. Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences*, 11(8):319–321, 2007.

[Pan and Slater, 2011] Xueni Pan and Mel Slater. Confronting a moral dilemma in virtual reality: a pilot study. In *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, pages 46–51. British Computer Society, 2011.

[Pedersen *et al.*, 2018] Bjarke Kristian Maigaard Kjær Pedersen, Kamilla Egedal Andersen, Simon Köslich, Bente Charlotte Weigelin, and Kati Kuusinen. Simulations and self-driving cars: A study of trust and consequences. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 205–206. ACM, 2018.

[Roberts and Torgerson, 1999] Chris Roberts and David J Torgerson. Baseline imbalance in randomised controlled trials. *Bmj*, 319(7203):185, 1999.

[Rovira *et al.*, 2009] Aitor Rovira, David Swapp, Bernhard Spanlang, and Mel Slater. The use of virtual reality in the study of people's responses to violent incidents. *Frontiers in Behavioral Neuroscience*, 3:59, 2009.

[Salem *et al.*, 2015] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 141–148. ACM, 2015.

[Shariff *et al.*, 2017] Azim Shariff, Jean François Bonnefon, and Iyad Rahwan. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour*, 1(10):694–696, 2017.

[Skitka *et al.*, 1999] Linda J Skitka, Kathleen L Mosier, and Mark Burdick. Does automation bias decision-making? *International Journal of Human-Computer Studies*, 51(5):991–1006, 1999.

[Theodorou *et al.*, 2017] Andreas Theodorou, Robert H. Wortham, and Joanna J. Bryson. Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science*, 29(3):230–241, 7 2017.

[Turkle, 2017] Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Hachette UK, 2017.

[Wortham *et al.*, 2017] Robert H Wortham, Andreas Theodorou, and Joanna J Bryson. Robot transparency: Improving understanding of intelligent behaviour for designers and users. In *Conference Towards Autonomous Robotic Systems*, pages 274–289. Springer, 2017.