Guilt without Fault: Accidental Agency in the Era of Autonomous Vehicles

Fernando Aguiar [a], Ivar R. Hannikainen[b], and Pilar Aguilar[c]*

[a]Institute of Philosophy, CSIC, Madrid, Spain; [b]Department of Philosophy, Universidad de Granada, Granada, Spain; [c]Department of Psychology, Universidad Loyola Andalucía, Sevilla, Spain.

*corresponding author: mpaguilar@uloyola.es

[**This paper is a preprint of an article published in *Science and Engineering Ethics* (2022) 28:11. https://doi.org/10.1007/s11948-022-00363-8. Please quote the published version**].

Abstract

The control principle implies that people should not feel guilt for outcomes beyond their control. Yet, the so-called 'agent and observer puzzles' in philosophy demonstrate that people waver in their commitment to the control principle when reflecting on accidental outcomes. In the context of car accidents involving conventional or autonomous vehicles (AVs), Study 1 established that judgments of responsibility are most strongly associated with expressions of guilt--over and above other negative emotions, such as sadness, remorse or anger. Studies 2 and 3 then confirmed that, while people generally endorse the control principle, and deny that occupants in an AV *should* feel guilt when involved in an accident, they nevertheless ascribe guilt to those same occupants. Study 3 also uncovered novel implications of the observer puzzle in the legal context: Passengers in an AV were seen as more legally liable than either passengers in a conventional vehicle, or even their drivers--especially when participants were prompted to reflect on the passengers' affective experience of guilt. Our findings document an important conflict--in the context of AV accidents--between people's prescriptive reasoning about responsibility and guilt on one hand, and their counter-normative experience of guilt on the other, with apparent implications for liability decisions.

Keywords: autonomous vehicles; moral responsibility; control principle; liability.

Introduction

In recent years, semi-autonomous vehicles (which offer drivers the option to alternate between an autonomous navigation mode and a driving mode) have taken to the streets in many different cities around the world. Anticipating the arrival of *fully* autonomous vehicles[1] (AVs), research suggests that AVs will provide numerous advantages over semi-autonomous and conventional vehicles: They will decrease road pollution (Riaz & Niazi, 2016) reduce traffic (Wu, Bayen & Mehta, 2018), and aid functionally diverse people, especially the visually impaired (Bennet et al., 2020). All of these contributions will foreseeably render AVs central to the future of human mobility (Bissel et al, 2018).

At the same time, the proliferation of AVs may increase the frequency with which these vehicles are involved in accidents. In light of this, AVs ought to be programmed to make morally defensible decisions when an accident can be foreseen, but not avoided. This practical demand has mustered the interest of moral philosophers and psychologists: Under what circumstances should AVs sacrifice one pedestrian to save five (Awad et al., 2018a)? Should an AV risk the lives of its *own* passengers for the greater good? Probing laypeople's intuitions on these matters, Bonnefon and colleagues (2016) concluded that people generally favor AVs that choose to save the larger number of lives (see also Wallach & Allen, 2009; Li et al., 2016). This conclusion was reinforced by a global online survey of almost 40 million trade-off decisions. The dominant preference for utilitarian reactions to inevitable accidents was qualified only in a handful of exceptional circumstances: e.g., AVs should not sacrifice passengers to save pedestrians, women to

[1] A fully autonomous vehicle has been defined by the United States of America's National Highway Traffic Safety Administration (NHTSA 2016) as the following: "The vehicle is designed to perform all safety-critical driving functions and monitor roadway conditions for an entire trip. Such a design anticipates that the driver will provide destination or navigation input but is not expected to be available for control at any time during the trip. This includes both occupied and unoccupied vehicles. By design, safe operation rests solely on the automated vehicle system". (NHTSA,2016).

save men, or children to save the elderly (Awad et al., 2018a, Gill, 2021). Additionally, people rejected self-sacrifice when they themselves were described as AV passengers, regardless of the number of occupants and pedestrians in danger (Bonnefon et al., 2016) [2]. These studies have since been met with some skepticism by researchers who denounce the reductionist and unrealistic nature of these trade-off scenarios (Nyholm & Smids, 2016; Santoni de Sio, 2017; Himmelreich, 2018; Rodríguez-Alcázar et al, 2020; Coeckelbergh, 2020) --inherited from the well-known trolley problem.

Guilt, Responsibility and Control

Alongside debates about the programming of AV guidelines, the development of AV technology has inspired further ethical questions. For instance, how is blame to be allocated when an AV runs over a pedestrian? Are passengers in an AV to be held morally responsible in the event of a crash? If so, on what grounds? These quandaries are of theoretical interest to ethicists, and have imminent practical implications for the AV industry and for the development of legislation on these matters (McManus et al, 2019; Gill, 2020). In the present paper, we explore these questions through the lens of a psychological puzzle (Struchiner et al., 2020): the conflict between people's abstract normative standards and their concrete experiences of guilt.

Existing research has explored people's intuitions regarding the allocation of blame in semi-autonomous vehicles. When a human driver and a semi-autonomous car share control over the vehicle, and both err, people ascribe greater blame to the human agent and exculpate the machine (Awad et al., 2018b). This tendency to direct blame toward human agents (and away from machines) extends to the manufacturer who

[2] It seems that Mercedes-Benz intended to advertise its AV by ensuring that they would prioritize the occupant's life above all: "Car and Driver reported on October 7 that Mercedes-Benz has already made a decision that its self-driving cars will always prioritize their own occupants over other road users. The automaker said it was wrong — and possibly illegal" (https://www.businessinsider.com/mercedes-denies-claim-its-driverless-car-will-prioritize-driver-safety-2016-10?IR=T, retrieved March 2, 2020).

programmed the AV as well as the lawmaker who enabled its circulation (Li et al., 2016; Pöllänen et al., 2020).

Meanwhile, in *fully* autonomous vehicles--the focus of our present studies--occupants play practically no role in the vehicle's driving decisions. Therefore, according to the *control principle*, i.e., the notion that an agent can only be morally responsible for actions over which they have control (Nagel, 1979; Fischer & Ravizza, 1998; Willemsen, 2019, Nelkin 2021), it stands to reason those passengers ought not be held responsible for the outcomes that ensue when riding an AV. Although they may have chosen to ride in the AV, this decision alone does not constitute the requisite control over any accidents that ensue (Pöllänen et al., 2020). Without this degree of control over the vehicle, passengers in an AV cannot--according to the control principle--be the objects of moral evaluation (DeRivera, 1984; Lindsay-Hartz, 1984; Bedford & Hwang, 2003; but see Bieker, 2012).

Yet Bernard Williams's (1981) *lorry driver* case casts doubt on this simple line of reasoning. In this thought experiment, a lorry driver is driving down the road when a child suddenly crosses the road in front of the lorry. Sadly, the driver can do nothing to avoid running over the child. Because there was no way the lorry driver could have avoided the tragic outcome, the driver lacked control. An impartial observer may recognize that, therefore, the driver cannot be held responsible for the accident, and even that he may justifiably eschew any feelings of guilt. At the same time, individuals who have lived through similarly tragic events attest to feeling guilty over their involvement (Gregory, 2017). This conflict has given rise to what are known as two related puzzles:

> First, why do these agents feel guilt (or something in the neighborhood of guilt) when they could hardly be held morally responsible for what happened? Are the agents' feelings appropriate even though the agents themselves aren't at fault or

> blameworthy? Second, in addition to this Agent Puzzle, there is also an Observer Puzzle: why do observers judge both that agents should not feel guilty and that if they do not feel guilty, they are deficient in some way? (Kamtekar & Nichols, 2019, p.2-3).

As we can see, both puzzles focus on the propriety of the agent's feelings of guilt: Agents may feel guilty and observers may view such a reaction as warranted in one sense, while both may recognize that guilt is inappropriate in another sense (Enoch 2012). According to Kamtekar and Nichols (2019), guilt is inappropriate because the action was not voluntary. For an action to be voluntary it must satisfy both (i) the aforementioned *control* condition, and (ii) an *epistemic* condition, i.e., the agent must know the circumstances in which the action takes place (Campos, 2013).

In cases of involuntary action, agents might reasonably feel various negative emotions: e.g., sadness about the consequences of the accident, frustration and regret over not being able to avoid it, and even anger at their bad luck. However, from a *rationalist* perspective, guilt is inappropriate in cases of involuntary action because the agent lacks the requisite control (Wolf, 2000). *Strawsonian approaches* offer a distinct analysis, but arrive at a similar conclusion (Strawson, 1962; Hieronymi, 2004; Carlsson, 2017). An agent can be treated as guilty when it is appropriate or just to punish them: i.e., through the expression of a reactive attitude (resentment, anger, or indignation) whose propositional content is correct. When agents are blameworthy, Strawsonian approaches claim that they deserve the sanction of others and ought to experience guilt (Nelkin, 2013; Clark, 2016; Carlsson, 2017). In the lorry driver case, however, if we view the driver as blameless, these approaches would lead to the conclusion that the lorry driver does not deserve the sanction of others' anger and indignation, or of their own experience of guilt.

Patricia Greenspan (1992) defends an alternative view of guilt without fault: Her *non-rationalist* ("non-judgmentalist") thesis is that in these cases agents act conditionally, i.e., *as if* they had moral responsibility. What is important here is not the propositional content of the guilt reaction (whether or not it conforms to a true or false belief about blameworthiness), but rather the agent's disposition to express "moral solidarity" (Greenspan, 1992, p. 300). In this regard, from the non-judgmentalist perspective, guilt may be appropriate--as a means to signaling moral solidarity under the pretense of conditional moral responsibility (see also Baumeister et al 1994; Deem & Ramsey, 2016).[3]

Philosophers have thus debated about the propriety of guilt in cases of accidental agency: Does the absence of control render feelings of guilt inappropriate? Some recent evidence indicates that AV passengers would indeed feel guilt when involved in an accident (Gill, 2020). What is not yet known is whether laypeople view guilt in these circumstances as appropriate or inappropriate. Our present work addresses this empirical question through three experimental studies. Answering this question could have important theoretical upshot for the debate between rationalist and non-rationalist perspectives on the problem of accidental agency.

By comparing AV passengers to the occupants of conventional human-driven vehicles, our studies may also yield practical insights. The proliferation of AV technology will demand legislation to govern instances in which automated vehicles are involved in traffic accidents. Our studies therefore probe people's judgments of responsibility and

[3] That the guilt feeling is appropriate does not mean that it is rational (as advocated by Sussman 2018), nor does it mean that it is irrational, as the logic of appropriateness is not necessarily linked to the logic of consequences (Balsige, 2014), which should be understood in this context as the (necessary) connection between the principle of control and the guilt feeling.

liability for accidents, to assess whether different intuitive standards are applied to accidents involving autonomous versus human-driven vehicles.

Current Research

To address these questions, in the current research we conducted three studies.[4] First, we explored which specific emotions AV occupants would feel in case of an accident and whether they differ from the emotions that a driver would feel after a conventional car accident (Study 1). Second, if AV passengers report feeling guilty, do they also deny that they *should* feel guilt? We examined this question through both the agent and observer perspective (Study 2). Finally, we evaluate participants' intuitions on the legal liability of AVs passengers in casualty accidents (Gill, 2020) (Study 3).

In sum, the current research was designed with five main objectives:

(1) Objective 1: To evaluate the predominance of guilt and other emotions among AV users and human drivers following a hypothetical accident.

(2) Objective 2: To evaluate the role of responsibility ascriptions in AV users' and human drivers' reports of both felt (descriptive) and prescriptive guilt following a hypothetical accident.

(3) Objective 3: To compare participants' assessments of guilt and responsibility following hypothetical accidents from the agent versus observer perspective (Kamtekar & Nichols, 2019).

(4) Objective 4: To compare participants' assessments of guilt and responsibility of both felt (descriptive) and prescriptive guilt between passengers of an AV and passengers of a conventional vehicle.

[4] The Ethics Committee of XXXX- has favorably evaluated this research (Committee Internal Code: 113/2020).

(5) Objective 5: To evaluate the downstream implications of AV technology and the phenomenon of in the legal context: i.e., on determinations of manslaughter for AV passengers, relative to drivers and passengers in conventional vehicles.

General Methods

Frequentist statistical analyses. were conducted using the *R* programming language. To examine the influence of our experimental factors, we conducted factorial ANOVAs (*analyses of variance*). We then followed up these analyses with tests of the simple and marginal effects of each factor, to discern the direction and magnitude of the effects. We report the magnitude of each effect in terms of Cohen's *d*. Study data may be found online on the *Open Science Framework* at: https://osf.io/8ch3x/?view_only=b5cd506c5ec4495898a03b558f3483d1.

Study 1

To assess Objective 1, we described a hypothetical car accident and examined the predominance of guilt by comparison to other negative emotions such as pity, remorse, sadness or anger (Greenspan, 1992; Wolf, 2000; Cordner, 2007; Katchadourian, 2010; Tangney et al., 2011; Nelissen, 2011, Carlsson, 2017) among AV passengers versus human drivers.

This perspective is consistent with various experimental findings and philosophical insights showing that remorse is a central component of guilt, a victim-oriented social emotion that only manifests itself interpersonally. Guilt would thus be the signal to others that harming them voluntarily or involuntarily causes a bad conscience or remorse, as well as the desire for reparation (Baumeister et al. 1994; Taylor, 1996; Cordner, 2007; Katchadourian, 2010; Nelissen, 2011; Tangney et al., 2011). This signal would be also a way of investing in moral reputation (being a better person) --people expect others to feel guilty in cases of blameless accidents (Rajen et al, 2021).

Method

Participants

The sample consisted of 200 United States adults recruited through Amazon Mechanical Turk, and was 46% female. Approximately half the sample was non-religious (49%), and the remaining half was largely Roman Catholic (20%) and Protestant (20%). Approximately half the sample were college graduates (49%), and the remaining half reported a lower educational attainment (e.g., complete secondary school, or some university education). The sample was politically diverse, with median political orientation being 'Moderate', and the interquartile range spanning from Q_1 = 'Liberal' to Q_3 = 'Conservative'. A signed-rank test confirmed that median political orientation did not significantly differ from 'Moderate', $V = 6150$, $p = .30$. Due to a technical issue, our study did not record participants' ages.

Procedure

At the beginning of the study, participants provided their informed consent. In a 2 x 2 between-subjects design, participants were randomly assigned to one of the following four conditions:

1. Autonomous Vehicle, Descriptive Condition.
2. Autonomous Vehicle, Prescriptive Condition.
3. Human Driver Car, Prescriptive Condition.
4. Human Driver Car, Descriptive Condition.

Participants were then asked to imagine an inevitable collision with a pedestrian, and to consider whether the passenger (Autonomous Vehicle condition) or driver (Driver condition), would be responsible for the accident. The responsibility measure was rated on a 7-point Likert scale from 1 (not responsible at all) to 7 (fully responsible).

On the following screen, participants viewed a list of five emotions (guilt, sadness, anger, pity and remorse) and were asked to report whether the driver/passenger 'would' or 'should' feel each emotion, in the Descriptive and Prescriptive conditions respectively. Emotions were rated on a 7-point Likert scale from 1 (not at all) to 7 (very much). In our primary analysis of emotional responses, we treat emotion as a within-subjects factor with five levels.

Results

Responsibility Ascriptions

In a two-way factorial ANOVA, we examined whether responsibility ascriptions depended upon vehicle-type and frame. We found a main effect of vehicle-type, $F_{(1,\ 196)} = 8.81$, $p < .003$, but effect of frame or two-way interaction, *p*s > .37. The marginal effect of vehicle-type indicated that passengers in AVs ($M = 3.92$, 95% CI [3.53, 4.32])) were ascribed less responsibility than were human drivers ($M = 4.72$, 95% CI [4.36, 5.08]), $t = -2.96$, $p = .004$.

Differences Between Emotions

Next, we 'stacked' the emotion variables resulting in a single continuous *intensity* rating, and a nominal variable which encoded the *emotion-type* (ie., guilt, remorse, sadness, anger and pity). We then conducted a three-way ANOVA with vehicle type and frame as between-subjects factors, emotion as a within-subjects factor (as well as the two- and three-way interactions). This model revealed a main effect of emotion, $F_{(4,\ 771)} = 36.56$, $p < .001$, which was qualified by an emotion×vehicle type interaction, $F_{(4,\ 771)} = 2.78$, $p = .026$. We observed no main effects of vehicle-type, $F_{(1,\ 193)} = 0.27$, or frame, $F_{(1,\ 193)} = 1.61$, $p = .21$, or interactions involving these variables, *F*s < 1, *p*s > .46.

Focusing on the descriptive ratings ("would feel"), we conducted a series of pairwise comparisons between emotions (i.e., of the main effect of emotion across vehicle-type) while applying Tukey correction for multiple comparisons. Participants reported feeling less anger than any other emotion (all *p*s < .001), and less pity than sadness (*p* = .006). Guilt and remorse ratings were non-significantly greater than pity ratings (*p*s > .13) and non-significantly lower than sadness ratings (*p*s > .52).

Turning to the normative ratings ("should feel"), corresponding analyses of the main effect of emotion (across vehicle-type) revealed the following pattern: Participants believed they *should* feel less anger than any other emotion (all *p*s < .005), and less pity than sadness (*p* = .005)--as in the descriptive condition. Guilt and remorse ratings were again non-significantly greater than pity ratings (*p*s > .13). This time, however, participants reported that they *should* feel less guilt than sadness (*p* = .012), though no less remorse than sadness (*p* = .78).

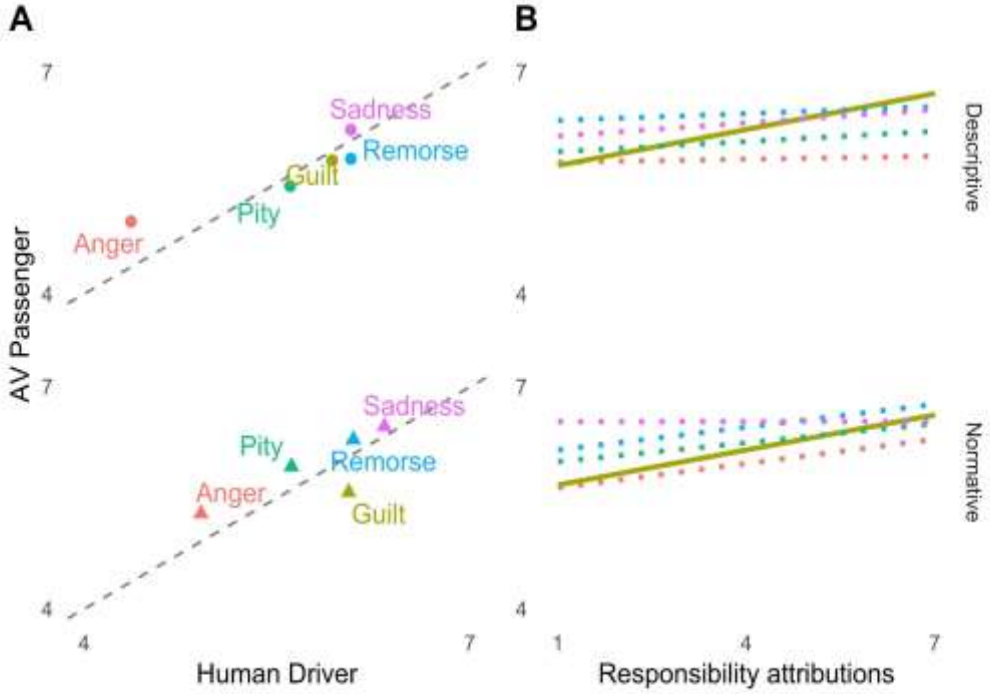

Figure 1. (A) Mean emotion intensity in the Human Driver (*x*-axis) and AV Passenger (*y*-axis) conditions. (B) Mean emotion intensity (*y*-axis; collapsed across vehicle-type) by responsibility ascriptions. Emotions obey the same color-coding in both plots. Responses in the Descriptive condition are displayed in the top pair of figures, and responses in the Normative condition are displayed in the bottom pair of figures

Thus, participants believed one would experience the most sadness, together with feelings of remorse and guilt. Then, when asked what emotions one should feel, participants endorsed feelings of sadness and remorse the most, and guilt to a slightly lower extent.

Associations Between Emotions and Responsibility

Next, we entered responsibility ascriptions into the ANOVA as a moderator. In a three-way ANOVA, we entered frame and responsibility ascriptions as between-subjects factors, and emotion as a within-subjects factor, allowing these terms to interact amongst each other. We included vehicle-type as a dummy code *only*, having observed no effects in our previous analysis and in order to reduce the complexity of the model.

This model revealed main effects of responsibility ascriptions, $F_{(1,\ 195)} = 30.99$, and emotion, $F_{(4,\ 783)} = 13.81$, which were qualified by a responsibility×emotion interaction, F(4, 783) = 7.09, all ps < .001. The two-way interaction indicated that the effect of responsibility ascriptions on emotion intensity varied across emotions. We therefore followed up on the two-way interaction by examining the pattern of simple slopes for each emotion.

In the descriptive condition, the simple slope of responsibility was steepest for guilt ($B = 0.45$, $p < .001$), followed by remorse ($B = 0.35$, $p < .001$), anger ($B = 0.22$, $p = .008$), and pity ($B = 0.17$, $p = .040$), and was non-significant for sadness ($B = 0.14$, $p = .088$). In the normative condition, the marginal slope of responsibility was once again steepest for guilt ($B = 0.38$, $p < .001$), followed by anger ($B = 0.27$, $p = .001$) and remorse ($B = 0.19$, $p = .020$), and was non-significant for pity ($B = 0.11$, $p = .17$) and sadness ($B = 0.06$, $p = .49$).

Discussion

Beliefs about responsibility best predicted feelings of guilt, even while other emotions (primarily sadness and remorse) would and should arise to a comparable degree.

Study 2

In Study 2, we focused on the role of responsibility judgments in AV passengers' and human drivers' reports of felt and prescribed guilt (Objective 2)--while setting aside all other emotion ratings. Additionally, to assess whether the agent and observer puzzles arise in the context of inevitable accidents, we evaluated the effect of participants' point of view (agent versus third party observer) on judgments of responsibility and guilt (Objective 3).

Method

Participants

The sample consisted of 201 United States adults recruited through Amazon Mechanical Turk, and was 49% female. Most participants were either non-religious (38%), Roman Catholic (38%) or Protestant (19%). Half the sample were college graduates (50%), and the remaining half reported a lower educational attainment (e.g., complete secondary school, or some university education). Once again, the sample was politically diverse: Median political orientation was non-significantly different from 'Moderate' in a signed rank test, $V = 6991$, $p = .71$, and the interquartile range spanning from Q_1 = 'Liberal' to Q_3 = 'Conservative'.

Procedure

In a 2 (vehicle-type: AV passenger, vs. human driver) x 2 (point-of-view: agent, vs. observer) x 2 (frame: descriptive, vs. prescriptive) between-subjects design, participants were randomly assigned to one of the following eight conditions:

1. AV passenger, Agent, Descriptive Condition;
2. AV passenger, Agent, Prescriptive Condition;
3. AV passenger, Observer, Descriptive Condition;
4. AV passenger, Observer, Prescriptive Condition;

5. Human Driver, Observer, Prescriptive Condition;
6. Human Driver, Agent, Prescriptive Condition;
7. Human Driver, Observer, Descriptive Condition;
8. Human Driver, Agent, Descriptive Condition;

In every condition, we described an inevitable accident involving a pedestrian's death narrated either in the first or the third person. As in Study 1, participants were asked whether the AV passenger (AV Passenger Condition) or the driver (Human Driver Condition) was responsible for the accident on a 7-point Likert scale, ranging from 1 (not responsible at all) to 7 (fully responsible). Participants were then asked whether the driver/passenger would (Descriptive Condition) or should (Prescriptive Condition) feel guilty in those circumstances, on a 7-point Likert scale, anchored at 1 (not guilty at all) and 7 (very guilty).

Finally, participants were asked whether they endorse a formulation of the control principle ("One should feel responsible/guilty for what is within one's control, but not for whatever is beyond one's control"), on a 7-point scale ranging from 1: 'Strongly disagree' to 7: 'Strongly agree'.

Results

Endorsement of the Control Principle

In a three-way factorial ANOVA with vehicle-type (AV passenger, vs. human driver), point-of-view (agent, vs. observer), and frame (descriptive, vs. prescriptive) as between-subjects factors, we observed no effects on whether participants endorsed the control principle, all *p*s > .29. We therefore collapsed across experimental conditions, and conducted a one sample *t*-test against the scale midpoint (mu = 4) to ascertain whether participants endorsed the control principle. We documented substantial agreement with the control principle ($M = 5.34$, $SD = 1.34$), $t = 10.31$, $p < 001$, Cohen's $d = 1.00$. In other words, participants believed in the abstract that agents should feel neither guilty nor responsible for outcomes beyond their control.

Responsibility Ascriptions

Second, in a three-way factorial ANOVA, we found a main effect of vehicle-type, $F_{(1, 193)} = 13.66$, $p < .001$. No significant effects of frame or person, *p*s > .20, or interaction effects, *p*s > .14, were observed in this model. The main effect of vehicle-type indicated that passengers in AVs (*M* = 3.56, 95% CI [3.18, 3.95])) were ascribed less responsibility than were human drivers (*M* = 4.60, 95% CI [4.21, 4.99]), *t* = -3.75, $p < .001$.

Guilt Ratings

Next, to assess the effect of our independent variables on participants' reports of guilt, we conducted a three-way ANOVA. Results revealed main effects of both vehicle type, $F_{(1, 193)} = 10.97$, $p = .001$, and frame, $F_{(1, 193} = 17.78$, $p < .001$. The main effect of vehicle type reflected significantly higher guilt ratings for human drivers (*M* = 5.57, 95% CI [5.21, 5.93]) than for AV passengers (*M* = 4.65, 95% CI [4.29, 5.00]), *t* = 3.60, $p < .001$. Meanwhile, the effect of frame revealed that descriptive guilt (*M* = 5.60, 95% CI [5.24, 5.96]) exceeded prescribed guilt (*M* = 4.61, 95% CI [4.26, 4.97]), *t* = 3.86, $p < .001$.

We did not observe a main effect of point of view, $F_{(1, 189)} = 0.33$, $p = .57$. In other words, adopting the agent versus the observer perspective did not influence guilt reports. Moreover, we did not observe any two- or three-way interactions between conditions, all *p*s > .05.

Associations Between Guilt and Responsibility

Our previous analyses indicated that vehicle-type affected judgments of responsibility, and reports of both descriptive and prescriptive guilt: Specifically, passengers in an AV were perceived as less responsible and also less guilty than human drivers, following a hypothetical accident.

Next, we entered responsibility ascriptions into the ANOVA as a (continuous) moderator of the main effects of frame and vehicle-type. Conceptually replicating Study 1, judgments of responsibility predicted guilt ratings, $F_{(1, 192)} = 105.34$, $p < .001$—though the effect was moderated by frame, $F_{(1, 192)} = 19.72$, $p < .001$. Specifically, in a pairwise contrast, the main effect of responsibility on prescriptive guilt, $B = 0.69$, 95% CI [0.55, 0.83], was larger than its effect on descriptive guilt, $B = 0.23$, 95% CI [0.07, 0.39], $t = 4.27$, $p < .001$ (see also Figure 2).

Discussion

Replicating Study 1, participants ascribed reduced responsibility to AV occupants than to drivers. Furthermore, as in Study 1, responsibility ascriptions were associated with guilt judgments.

Additionally, we observed that drivers and AV passengers alike *would* feel guilt, but that they should not feel guilt to the same degree. In other words, though participants acknowledged that AV occupants and drivers would experience guilt, they tended to deny that such guilt would be normatively appropriate in the absence of control over the vehicle. This result dovetails with the finding that participants tended to endorse the control principle in the abstract.

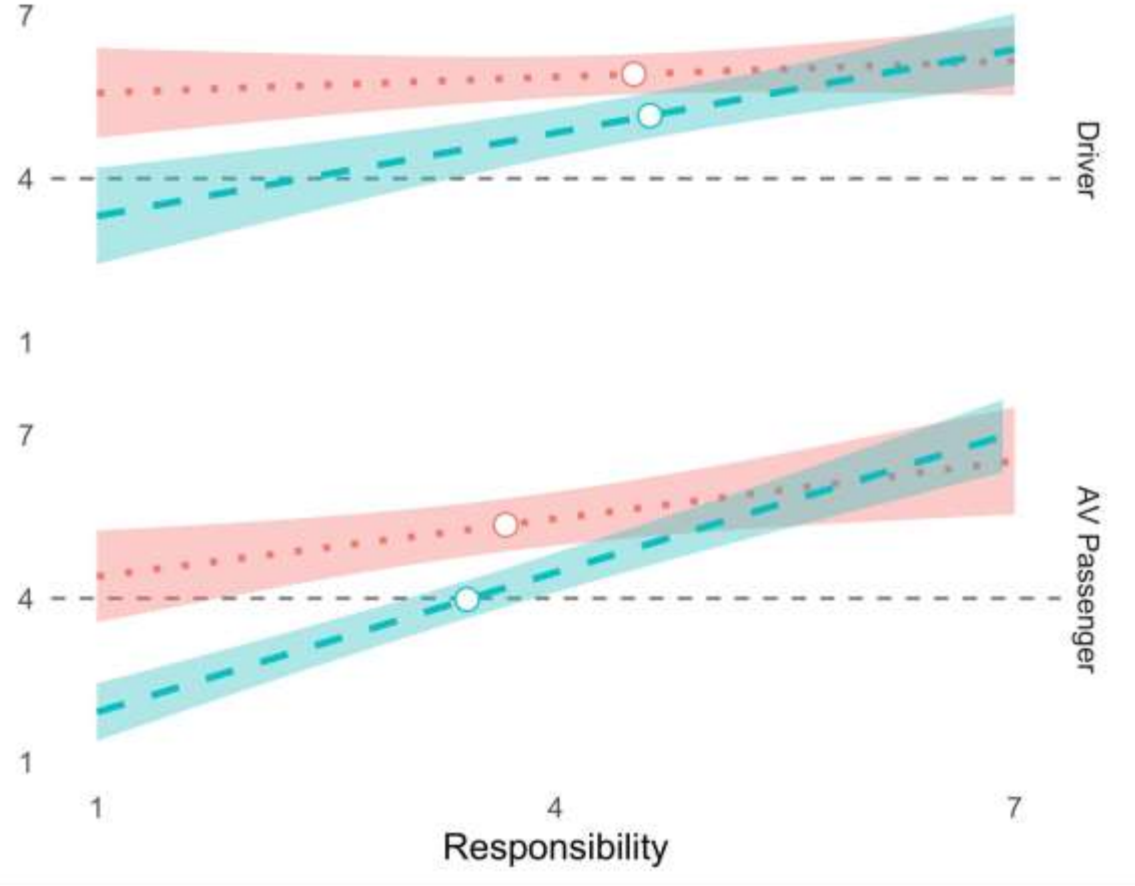

Figure 2. Effects of responsibility on prescriptive (blue) and descriptive (red) guilt in the Driver (top) and AV Passenger (bottom) conditions. Overlaid points represent the condition means.

Study 3

We obtained mixed evidence of whether AVs reduce descriptive and prescriptive guilt relative to conventional vehicles: In Study 2, we observed an effect that was absent in Study 1. So, we conducted a pre-registered replication with a larger sample size to disambiguate this result in Study 3.

To this end, we added a second control condition in order to match the protagonists' passenger status, by comparing a passenger in an AV to a passenger in a conventional vehicle (Objective 4). Finally, to assess the legal implications of the adoption of AV technology, we asked participants two further questions. First, participants decided whether protagonists would be liable for involuntary manslaughter--after learning that the pedestrian had died in the hospital. Second, they were asked to impose an appropriate sentence on the protagonist (Objective 5).

Method

Participants

We recruited 604 native English speakers through Prolific Academic, of whom 92 participants failed a comprehension question at the end of the study. and were excluded from the analyses. The remaining sample (N = 514) was made up of 68% women. Ages ranged from 19 to 75, with the median age being 32 years old. The dominant nationality was the United Kingdom (71%), followed by South Africa (9%) and Canada (7%).

Procedure

Participants were randomly assigned to one of the following six conditions in a 3 (protagonist: AV passenger, conventional driver, conventional passenger) x 2 (frame: normative, descriptive) between-subjects design:

1. AV Passenger, Descriptive Condition;
2. AV Passenger Prescriptive Condition;
3. Conventional Driver, Descriptive Condition;
4. Conventional Driver, Prescriptive Condition;
5. Conventional Passenger, Descriptive Condition;
6. Conventional Passenger, Prescriptive Condition.

In Study 2 we did not find any effect of participants' point of view so we abandoned this manipulation and narrated every scenario from the third person perspective. In every condition, we described an inevitable accident involving a pedestrian's death, and displayed a realistic photograph of a car accident site to render the scenario more evocative. Participants were then asked whether the protagonist (either the driver, or the passenger in an autonomous or conventional vehicle) would (Descriptive condition) or should (Prescriptive condition) feel guilty in those circumstances on a 100-point Likert scale, anchored at 0 (not guilty at all) and 10 (very guilty), in 0.1 increments.

Additionally, participants were asked two questions to examine the legal implications of our manipulations of (i) the protagonist's role and (ii) the descriptive versus prescriptive framing of the emotion appraisal.

Participants were told "Now suppose that the pedestrian, having sustained serious injuries due to the collision, dies in the hospital several hours later". They then rated the protagonist's *liability* ("Based on what you know about the accident, should Jordan be charged with involuntary manslaughter?") on a 10-point Likert scale, anchored at 0 (certainly not) and 100 (certainly), in 0.1 increments. Next, participants made sentencing decisions ("The charge of involuntary manslaughter ranges from no prison time (community service and/or probation) to 12 years of prison time, depending on the circumstances. Suppose you are serving on a jury, what charges would you recommend for Jordan?") on a 13-point scale ranging from 0: 'no jail time (probation and/or community service)' to 12: '12 years (maximum prison time)'. Lastly, as in Study 2, participants reported whether they agree or disagree with the control principle on a 7-point scale.

Pre-registration

Sample size, design and analysis plan were pre-registered at: https://aspredicted.org/TUJ_NZM.

Results

Endorsement of the Control Principle

In a 3 (protagonist) x 2 (frame) ANOVA, we found no effects on whether participants endorsed the control principle, all *p*s > .27. We therefore collapsed across experimental conditions, and conducted a one sample *t*-test against the scale midpoint (mu = 4). Replicating Study 2, we found general agreement with the control principle (*M* = 5.33, *SD* = 1.38), *t* = 21.80, *p* < 001, Cohen's *d* = 0.96.

Guilt Ratings

Next, we conducted a 3 (protagonist) x 2 (frame) ANOVA entering guilt ratings as the dependent measure. This model revealed main effects of protagonist, $F_{(2,\,506)}$ = 29.71, and frame, $F_{(1,\,506)}$ = 151.50, both *p*s < .001, but no two-way interaction, $F_{(2,\,506)}$ = 1.27, *p* = .28.

To examine the main effect of protagonist, we conducted all three pairwise comparisons and adjusted the resulting *p*-values using Tukey correction. Participants provided higher guilt ratings for conventional drivers (*M* = 6.96, 95% CI [6.61, 7.30]), *t* = 6.99, and AV passengers (*M* = 6.86, 95% CI [6.49, 7.24]), *t* = 6.45, than for passengers in conventional vehicles (*M* = 5.06, 95% CI [4.65, 5.46]), both *p*s < .001. In contrast, guilt among drivers and AV passengers did not differ, *t* = 0.36, *p* = .93.

Meanwhile, the effect of guilt-type (averaged over levels of protagonist) revealed that descriptive guilt (*M* = 7.63, 95% CI [7.32, 7.93]) exceeded prescriptive guilt (*M* = 4.96, 95% CI [4.65, 5.26]), *t* = 12.12, *p* < .001. This result provided a pre-registered replication of Study 2.

Liability for Involuntary Manslaughter

Second, we conducted the corresponding 3 x 2 ANOVA with liability for involuntary manslaughter as the dependent measure. Results revealed main effects of protagonist, $F_{(2,\,506)}$ = 39.00, and frame, $F_{(1,\,506)}$ = 11.14, *p*s < .001, but no two-way interaction, $F_{(2,\,506)}$ = 0.03, *p* = .97.

A series of pairwise comparisons indicated that participants viewed drivers (*M* = 3.57, 95% CI [3.20, 3.94]) as more liable than passengers (*M* = 2.26, 95% CI [1.83, 2.70]) in a conventional vehicle, $t = 4.49$, $p < .001$. Strikingly, AV passengers' liability (M = 4.90, 95% CI [4.51, 5.30]) was significantly *greater* than the liability of either drivers or passengers in a conventional vehicle, both ts > 4.81, ps < .001.

Meanwhile, the effect of frame revealed greater liability after considering whether the protagonist *would* feel guilty (M = 3.97, 95% CI [3.65, 4.30]) than after assessing whether they *should* feel guilty (M = 3.19, 95% CI [2.86, 3.52]), $t = 3.31$, $p = .001$.

Sentencing Recommendations

A 3 x 2 ANOVA with sentencing decisions as the dependent measure revealed a main effect of protagonist, $F_{(2, 506)} = 15.93$, $p < .001$, but no effect of frame, $F_{(1, 506)} = 0.19$, $p = .67$, or of the two-way interaction, $F_{(2, 506)} = 2.81$, $p = .061$.

The protagonist effect indicated harsher sentences on drivers (M = 1.94, 95% CI [1.63, 2.25]) than on passengers in a conventional vehicle (M = 1.35, 95% CI [0.99, 1.71]), $t = 2.44$, $p = .040$. Once again, as with liability, sentencing decisions were harsher for AV passengers (M = 2.78, 95% CI [2.45, 3.12]) than for either drivers or passengers in conventional vehicles, both ts > 3.65, ps < .001.

Discussion

Our pre-registered replication in Study 3 confirmed that descriptive guilt exceeded ratings of prescriptive guilt: Protagonists should not feel guilt to the extent that they do after an inevitable accident. We observed no reduction in guilt for AV passengers relative to drivers (replicating Study 1), despite participants' widespread endorsement of the control principle.

Study 3 also uncovered downstream implications of both experimental manipulations in the legal domain. First, AV passengers' liability for accidents was greater than that of either passengers or drivers in conventional vehicles--a pattern that was mirrored in participants' sentencing recommendations.

Second, a descriptive frame--i.e., focusing on whether protagonists *would* feel guilt--elevated liability relative to a prescriptive frame--i.e., focusing on whether protagonists *should* feel guilt (see Fig. 3).

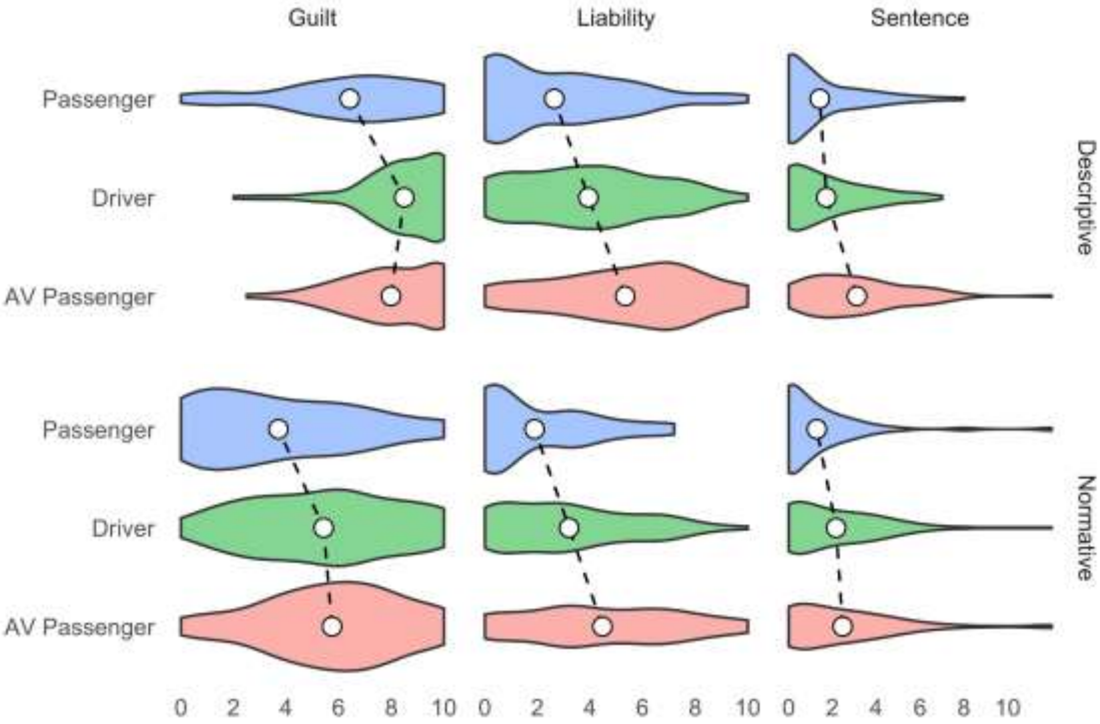

Figure 3. Violin plots of guilt, liability and sentencing recommendations for each protagonist. The overlaid dots represent the condition means.

General Discussion

Across three studies we examined whether the agent and observer puzzles arise in the context of autonomous vehicles accidents, as well as their downstream effects on liability in the legal domain. In the first study, responsibility ascriptions were more strongly associated with reports of guilt than with reports of other emotions. This held true even though reports of sadness were significantly greater than reports of guilt or remorse (Objective 1). Then, in Study 2, we found that participants tended to ascribe feelings of guilt, while denying that guilt would be normatively appropriate. Furthermore, mediation analyses indicated that attributions of responsibility better predicted prescriptive guilt (than descriptive guilt), pointing to a discrepancy between perceived responsibility and feelings of guilt (Objective 2). This pattern of results emerged equally whether participants adopted the agent's or an observer's perspective in an inevitable accident (Objective 3). Study 3 provided a pre-registered replication of the discrepancy between descriptive and prescriptive guilt, with the former exceeding the latter--as in Study 2. Turning to legal implications, we found that AVs passengers were dealt harsher sentences and greater liability than the occupants of conventional vehicles (Objective 4), including their drivers. Finally, we observed downstream effects of a descriptive versus prescriptive frame on judgments of liability, but not sentencing decisions (Objective 5).

To further distill the implications of our results, let us return to the distinction—following Kamtekar and Nichols, who in turn rely on Aristotle—between forced, involuntary and non-voluntary actions (Kamtekar & Nichols, 2019, see also Campos, 2013). We have already seen that an action is voluntary when it meets both the control and epistemic conditions. An action is involuntary, then, when it is forced (the person, as an inert body, does not contribute at all to the action) or when the epistemic condition is not fulfilled (the circumstantial facts are unknown) and, in such a case, "the agent feels pain or regret" (Kamtekar & Nichols, 2019, p. 186-187). Thus, when the action is involuntary but not forced there is a mismatch between intention and action "with negative evaluation" (p.187) This mismatch also occurs in the case of non-voluntary actions, although this subset of actions is not linked to negative emotions (Kamtekar & Nichols, 2019). Hence, although the "proper domain" of guilt concerns voluntary actions, its "current domain" includes involuntary ones, that is, cases of accidental agency (Kamtekar & Nichols, 2019, p. 192).

However, let's now imagine two passengers in an AV. One of them reads and the other sleeps. Suddenly they hear a loud thump, the AV comes to a halt, and a person lies severely injured on the roadway. The data from the onboard computer shows that the AV was unable to do anything to prevent the accident because the person suddenly crossed the road. What kind of agents are these passengers and what kind of action did they undertake? The accidental collision with the pedestrian is not voluntary, of course—since neither the control nor the epistemic conditions are met. It is not non-voluntary either, as it is linked—as our data shows—with negative emotions. We might therefore consider it an *involuntary non-forced action*, one in which there is no mismatch between intention and action, since their intention is largely absent: It is limited to the initial decision to travel in an AV, to choose the route, and to read or sleep along the way. AV passengers are, therefore, neither casually nor morally responsible for what happened, since the agentive nature of their actions is unrelated to the accident. Thus, they can hardly be considered agents in relation to the facts, cannot be blamed for the accident, and should not feel guilt—at most, sorrow over what happened. However, the participants claim that, alongside their sadness, they would also feel remorse and guilt (Study 1), and hold AVs passengers legally liable when involved in fatal accidents (Study 3). Why does this happen?

First, it could be thought that feeling guilt in AV accidents is not only inappropriate, but clearly irrational, so it should be excluded from the proper and current guilt domains (Nagel, 1979; Jacobson, 2012). Given that today most of us have yet to ride in an AV, we may wonder whether participants in our study were simply extrapolating from their experience with normal cars. However, we suspect this is not the case, since participants in our studies were at least cognizant of differences between AVs and conventional cars–for instance, by considering that AV passengers are legally more liable than passengers in conventional cars.

On the other hand, AV passengers do not deserve to be sanctioned by reactive emotions such as indignation or resentment. It would be odd for observers to express outrage at the occupants of an AV that had run over a pedestrian. From a Strawsonian perspective, it can be said that these passengers should not feel guilty (it is not fair to demand it) because they do not deserve to be blamed.

Still the agent and observer puzzles, documented in our studies, call for an explanation. Potentially, guilt, together with sadness and remorse, serves as public signs that we regret what happened, that we identify with the victims (Greenspan, 1992). Failing to do so could also affect the moral reputation of those involved (Rajen et al, 2021). In other words, AVs passengers would behave, in Greenspan's terms, "as if" they were morally responsible for the accident, in which case their subjective or felt guilt would be understandable, even in the absence of fault. Indeed, this could be a way to understand why participants reported higher guilt ratings for drivers and AV passengers than for passengers in conventional vehicles (see Study 3). In contrast to passengers in a conventional vehicle, participants seem to consider that the AV passengers would have a comparable moral status, in guilt terms, to that of drivers in conventional cars. This may also explain why there is no reduction in guilt for AV passengers relative to drivers, despite participants' widespread endorsement of the control principle: Guilt could in both cases represent an agent's display of "moral solidarity" (Greenspan, 1992, p. 300).

Greenspan's "as if" thesis might also help to understand why AV passengers in Study 3 are held *more* legally liable than passengers in conventional vehicles, again because of a similar moral status between drivers and AV passengers that would lead them to act as if they were morally responsible. However, it is not clear why it is also reported that AV passengers should be more legally liable than drivers.

Thus, the rationalist perspective, the Strawsonian perspective, and Greenspan's nonjudgmental view face different problems in accounting for people's guilt feeling in AVs accidents –*the AV puzzle*-, as well as in explaining the extent of the moral and legal consequences thereof.

Limitations and Future Research

The present work takes a step forward in the study of the moral implications of AVs, but it is not without limitations that need to be acknowledged. First, our studies were implemented online with convenience samples recruited through crowdsourcing services, such as Amazon Mechanical Turk. Despite the popularity of these recruitment methods (see e.g., Nguyen, Tina et al., 2019), concerns about low data quality have been put forth (see Huff & Tingley, 2015). To alleviate this concern, we ran our third study on Prolific.co, a similar service which has been shown to yield higher data quality than its competitors (Peer at al., 2021). In future studies should, nonetheless, seek to replicate these findings among more diverse, i.e., nationally representative, samples.

Second, in Study 3, we observed a counterintuitive finding: namely, that AV passengers were seen as *more* liable and awarded more severe sentences than drivers in conventional cars. This result may reflect the prevailing distrust that these technologies muster[5], and should be explored further. Relatedly, future research could employ an individual differences approach (e.g., examining neophobia, or openness to experience), to better understand the mechanism that underlies guilt and responsibility attributions in the context of autonomous vehicles. However, as our experimental studies apply random

[5] "AAA's [American Automobile Association] annual automated vehicle survey found that 71 percent of people are afraid to ride in fully self-driving vehicles" (Edmons, E., 2019). See also Xu et al, 2018.

assignment to conditions, we do not expect such factors to conflate the primary results reported in this work.

Third, the use of hypothetical vignettes, as the ones used in the current studies, has reported several limitations: most notably, that participants may respond, not as they would if they experienced the situation in real life, but based on their expectations of what constitutes a socially desirable answer (Lerner, 2013). We encourage future studies to adopt more realistic stimuli in order to faithfully simulate the experience of AV usage.

Conclusion

Ascribing moral and legal liability in the wake of accidental, and even inevitable, outcomes poses a complex theoretical and practical problem. In this paper, we examined how people resolve this conflict in the context of autonomous vehicle technology.

Taken together, our studies illustrate a certain continuity in the way we ascribe responsibility to passengers in autonomous vehicles and drivers in conventional vehicles. In the abstract, people overwhelmingly endorse the principle that agents should deflect any responsibility for outcomes that are beyond their control. Relatedly, people view expressions of guilt (more so than other negative emotions, such as sadness, pity or remorse) as displays of the admission of responsibility. As such, when considering an inevitable road traffic accident, people also tend to deny that passengers in autonomous vehicles should express feelings of guilt.

At the same time, they recognized that passengers in an autonomous vehicle *would* experience guilt in these same circumstances. The affective forecast of guilt had downstream consequences on people's liability decisions, resulting in a greater tendency to convict passengers and drivers of involuntary manslaughter. In sum, the 'counter-normative' experience of guilt could interfere with people's prescriptive reasoning about responsibility and control–with important implications for decision-making in the courtroom.

References

Awad, E., Dsouza, S., Kim, R., Schultz, J., Henrich, J., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2018a). The Moral Machine experiment. *Nature*, *563*, 59–64.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J-F., & Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour*, *4*(2), 134-143.

Balsiger, J. (2014). Logic of appropriateness. *Encyclopedia Britannica*. Available at http://www.britannica.com/EBchecked/topic/1937900/logic-ofappropriateness. Retrieved May 12, 2021.

Baumeister, R., Stillwell, A. & Heatherton, T. (1994). Guilt: An interpersonal approach. *Psychological Bulletin, 115 (2)*, 243-267.

Bieker, S. (2012). Legal aspects of autonomous driving. *Santa Clara Law Review, 52*, 1146-1156.

Bennet, R., Vijaygopal, R., & Kottasz, R. (2020). Willingness of people who are blind to accept autonomous vehicles: An empirical investigation. *Transportation Research Part F: Traffic Psychology and Behaviour*, *69*, 13-27.

Bissel, D., Birtchnell, T., Elliott, A., & Hsu, E. L. (2018). Autonomous automobilities: The social impact of driverless vehicles. *Current Sociology, 68*, 116–134.

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

Campos, A. S. (2013). Responsibility and justice in Aristotle's non-voluntary and mixed actions. *Journal of Ancient Philosophy*, *7*, 100-121.

Carlsson, A. B. (2017). Blameworthiness as deserved guilt. *Journal of Ethics, 21,* 89-115.

Coeckelbergh, M. (2020). *AI Ethics*. Cambridge MA: Mit Press.

Cordner, C. (2007). Guilt, remorse and victims. *Philosophical Investigations, 30*, 337-362.

Deem, M. J., & Ramsey, G. (2016). Guilt by association? *Philosophical Psychology*, *4*, 570-585.

Edmons, E. (2019). Self-driving vehicles. https://newsroom.aaa.com/2019/03/americans-fear-self-driving-cars-survey/. Retrieved 5 July, 2021.

Enoch, D. (2012). Being responsible, taking responsibility, and penumbral agency. In Heuer, U. & Lang, G., editors. *Luck, Value, and Commitment - Themes from the Ethics of Bernard Williams*. Oxford: Oxford University Press, pp. 95-132.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research*, *47* (2), 272–291,

Gill, T. (2021). Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics and Information Technology*, https://link.springer.com/article/10.1007/s10676-021-09605-y.

Greenspan, P. S. (1992). Subjective guilt and responsibility. *Mind, 101*, 287-303.

Gregory, A. (2017). The sorrow and the shame of the accidental killer. *The New Yorker*, September 18.

Hieronymi, P. (2004). The force and fairness of blame. *Philosophical Perspectives, 18,* 115-148.

Himmelreich, J. (2028). Never mind the trolley: The ethics of autonomous vehicles in mundane situations. *Ethical Theory and Moral Practice, 21,* 669–684.

Huff, C., & Tingley, D. (2015). 'Who are These People?′ Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents. *Research & Politics, 2*(3), 1–12.

Jacobson, D. (2012). Moral dumbfounding and moral stupefaction. *Oxford Studies in Normative Ethics, (2)*, 289-316.

Kamtekar, R., & Nichols, S. (2019). Agent-regret and accidental agency. *Midwest Studies in Philosophy*, *43*, 181-202.

Katchadourian, H. (2010). *Guilt. The Bite of Conscience*. Stanford: Stanford University Press.

Lerner, M. J. (2003). The justice motive: Where social psychologists found it, how they lost it, and why they may not find it again. *Personality and social psychology review*, *7*(4), 388-399.

Li, J., Zhao, X., Cho, M., Ju, W. & Malle, B. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *SAE Technical Paper* 2016-01-0164, doi:10.4271/2016-01-0164.

McManus, R. M. & Rutchick, A.M. (2019). Autonomous vehicles and the attribution of moral responsibility. *Social Psychological and Personality Science, 10,* 345-352.

Nagel, T. (1979). *Mortal Questions*. New York: Cambridge University.

Nelissen, R. (2011). Guilt-induced self-punishment as a sign of remorse. *Social Psychological and Personality Science, 3(2)*, 139-144.

Nelkin, D. (2013). Accountability and desert. *The Journal of Ethics 20(1),* 173–189.

Nelkin, D. K., (2021). Moral Luck. *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2021/entries/moral-luck/.

Nguyen, T., et al (2019). Metamotivational knowledge of the role of high-level and low-level construal in goal-relevant task performance. *Journal of personality and social psychology, 117 (5),* 876-899.

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

Peer, E., Rothschild, D. M. and Evernden, Z., Gordon, A., & Damer, E. (2021). MTurk, Prolific or Panels? Choosing the Right Audience for Online Research. Available at SSRN: https://ssrn.com/abstract=3765448

Pöllänen, E., Read, G., Lane, B. R., Thompson, J., & Salmon, P. M. (2020). Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system, *Ergonomics,* Advance online publication, DOI: 10.1080/00140139.2020.1744064.

Rajen A. A., Kamtekar, R., Nichols, Sh., & Pizarro, D. (2021). "False positive" emotions, responsibility, and moral character. *Cognition*, 214, https://doi.org/10.1016/j.cognition.2021.104770.

Riaz, F., & Niazi, M. A. (2016). Road collisions avoidance using vehicular cyber-physical systems: a taxonomy and review. *Complex Adaptive Systems Modeling*, *4*, 1-34.

Rodríguez-Alcázar, J., Bermejo-Luque, L. & Molina-Pérez, A. (2020). Do automated vehicles face moral dilemmas? A Plea for a Political Approach. *Philosophy Technology* https://doi.org/10.1007/s13347-020-00432-5.

Santoni de Sio, F. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice 20 (2)*, 411-429.

Struchiner, N., Almeida, G. F. C. F., & Hannikainen, I. R. (2020). Legal decision-making and the abstract/concrete paradox. *Cognition*, *205*, 104421.

Sussman, D. (2018). Is agent-regret rational? *Ethics*, *128*, 788-808.

Tangney, J. P., Stuewig, J. & Mashek, D. B. (2007). Moral Emotions and moral Behavior. *Annual Review of Psychology*, *58* (1), 345-72.

Taylor, G. (1996). Guilt and remorse. In R. Harre & W. Gerrod Parrot (eds.), *The Emotions: Social, Cultural and Biological Dimensions.* Sage Publications, pp. 57-73.

Taylor, G. (1985). *Pride, Shame and Guilt.* Oxford, UK: Oxford University Press.

Wallach, W. & C. Allen (2009). *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.

Willemsen, P. (2019). *Omissions and their Moral Relevance. Assessing Causal and Moral Responsibility for the Things we Fail to Do*. Münster: Mentis Verlag.

Williams, B. (1981). *Moral Luck.* Cambridge: Cambridge University Press.

Wolf, S. (2000). The moral of moral luck. *Philosophical Exchange, 31 (1)*, 2-19.

Wu, C., Bayen, A. M., & Mehta, A. (2018). Stabilizing traffic with autonomous vehicles. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1-7). IEEE.

Xu, Z., Zhang, K., Min, H., Wang, Z., Zhao, X., & Liu, P. (2018). What drives people to accept automated vehicles? Findings from a field experiment. *Transportation research part C: emerging technologies*, *95*, 320-334.

Appendix 1: Internal Meta-Analysis

Given the conflicting results, we conducted an internal meta-analysis (n = 768) to understand whether:

(1) perceived guilt feelings in the Descriptive frame exceed prescribed guilt in the Normative frame, and

(2) AV passengers are ascribed less guilt than conventional drivers.

Given differences in the scale length across studies, we first min-max normalized the dependent variable (i.e., from 0 to 1). A two-way ANOVA indicated that the effects of frame, $F_{(1, 764)} = 80.61$, $p < .001$, eta$^2_\mathrm{p}$ = .11, and vehicle-type, $F_{(1, 764)} = 9.01$, $p = .003$, eta$^2_\mathrm{p}$ = .01, were significant. The interaction was non-significant, $F_{(1, 764)} = 0.01$, $p = .92$. Thus, though both effects were significant, the effect of frame was medium to large, whereas the effect of vehicle-type was small.

Appendix 2: Gender Differences in Guilt

We also examined whether there are gender differences in guilt ratings. To do this, we entered gender as a third factor, allowing it to interact with both experimental manipulations. We observed a main effect of gender, $F_{(1, 757)} = 13.33$, $p < .001$, and a marginally significant interaction with frame, $F_{(1, 757)} = 2.82$, $p = .088$. The effect of frame was larger among women, $B = 0.21$, $t = 7.90$, than among men, $B = 0.14$, $t = 4.48$, ps < .001. This difference was due to higher guilt ratings among women ($M = 0.85$) than among men ($M = 0.74$) in the Descriptive condition, $t = 3.85$, $p < .001$. The corresponding effect in the Normative condition was non-significant ($M_{\text{-Women}} = 0.65$, $M_{\text{-Men}} = 0.61$), $t = 1.47$, $p = .14$.

Acknowledgements.

We would like to thank the participants in the experimental-ethics session at the *XIX Ethics and Political Philosophy Week,* Ferrol (Spain), October 2018, particularly Carlos Thiebaut, Blanca Rodriguez and Mabel Holgado, for very helpful feedback. We would also like to thank Mabel Holgado for making us pay special attention to the relationship between guilt and remorse.

This article is based on research funded by the Spanish Ministry of Science, Innovation, and Universities, reference number RTI2018-098882-B-I00 and reference number PID2020.119791RA.I00

Declarations

Conflicts of interest All procedures performed in the studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent Informed consent was obtained from all individual participants included in the study.

Ethical approval. Ethical approval was given by the Consejo Superior de Investigaciones Cientificas –Spanish National Research Council- (Committee Internal Code: 113/2020).