Improving the Public's Perception of Autonomous Vehicles by Communicating the Consistency of Autonomous Vehicle Algorithms

Heather E. K. Walker, M.Sc.
University of Guelph

Alexander C. Walker, M.A.
University of Waterloo

Rafał Muda, M.A.
Maria Curie-Skłodowska University

Martin Harry Turpin, M.A.
University of Waterloo

Lana M. Trick, Ph.D.
University of Guelph

Jonathan A. Fugelsang, Ph.D.
University of Waterloo

Michał Białek, Ph.D.
University of Wrocław

Abstract

Despite autonomous vehicles (AVs) being safer than human drivers, people are averse to their presence on roads. Across three studies (N = 4,014), we examined peoples' perceptions of human drivers and AVs acting within a moral dilemma. Scenarios involved an out-of-control vehicle (piloted by a human, or autonomously) that could stay on its present course and hit five pedestrians, or swerve and hit a single stranger. Participants were given a description of the pilot's final action and then judged them on several dimensions (e.g., blame, acceptability, predictability). We find evidence of AV aversion across all studies, with participants judging AVs more negatively (e.g., more blameworthy) than human drivers despite performing identical actions. Additionally, Studies 1 and 2 presented some participants with a statement outlining the consistency of AV algorithms, which increased perceived predictability and reduced AV aversion in some cases. In Study 3, some participants were given scenarios in which control of the vehicle was transferred prior to a pilot's actions. Participants' were averse to this transfer, as both AVs and human drivers were perceived as less predictable and judged more negatively after taking control of the vehicle. Overall, our findings highlight peoples' aversion to autonomous and semi-autonomous vehicles,

while also demonstrating that messages highlighting the consistency of AV algorithms have the potential to improve perceptions and thus reduce barriers for their eventual mass adoption.

Résumé

Bien que les véhicules autonomes (VA) soient plus sûrs que les conducteurs humains, les gens sont opposés à leur présence sur les routes. Dans trois études (N = 4 014), nous avons examiné les perceptions des gens sur les conducteurs humains et les VA agissant dans un dilemme moral. Les scénarios impliquaient un véhicule incontrôlable (piloté par un humain ou de manière autonome) qui pouvait rester sur sa trajectoire actuelle et heurter cinq piétons, ou dévier et frapper un seul étranger. Les participants ont reçu une description de l'action finale du pilote, puis les ont jugés sur plusieurs dimensions (par exemple, blâme, acceptabilité, prévisibilité). Nous trouvons des preuves d'un biais contre les VA dans toutes les études, les participants jugeant les VA plus négativement (par exemple, plus blâmables) que les conducteurs humains, malgré des actions identiques. De plus, les études 1 et 2 ont présenté à certains participants une explication décrivant la cohérence des algorithmes de VA, qui augmentaient la prévisibilité perçue et réduisaient le biais contre les VA dans certains cas. Dans l'étude 3, certains participants ont eu des scénarios dans lesquels le contrôle du véhicule était transféré avant les actions d'un pilote. Les participants étaient opposés à ce transfert, car les VA et les conducteurs humains étaient perçus comme moins prévisibles et jugés plus négativement après avoir pris le contrôle du véhicule. Dans l'ensemble, nos résultats mettent en évidence l'aversion des gens pour les véhicules autonomes et semi-autonomes, tout en démontrant également que les messages mettant en évidence la cohérence des algorithmes de VA ont le potentiel d'améliorer les perceptions et ainsi de réduire les obstacles à leur éventuelle adoption massive.

INTRODUCTION

According to the National Highway Traffic Safety Administration, approximately 94% of collisions in the US are caused, at least in part, by human error [1]. These collisions have a large impact on society, resulting in more than 36,000 deaths, $57 billion lost in workplace productivity, and $590 billion in compensation for injury or loss of life each year [1]. Autonomous vehicles (AVs) represent a promising avenue for improving roadway safety, as the wide-spread adoption of AVs is projected to reduce the frequency and severity of motor vehicle collisions by up to 90% [2–3]. Already, evidence from existing pilot projects show that AVs are involved in nearly half the number of collisions (per million miles driven) than human-piloted vehicles [4–5].

Despite the promising safety implications of AVs, people appear to be averse to their adoption. For instance, following several fatal incidents in 2018 involving vehicles operating in autonomous mode, distrust in AVs increased [6–8]. Approximately 71% of individuals surveyed shortly after indicated that they were afraid to ride in an AV—an increase from 63% prior to these events [7]. Consistent with this finding, Liu and colleagues [9] observed that AVs would need to be nearly five times safer than human-piloted vehicles in order to be considered acceptable. Peoples' aversion to AVs has even resulted in violence, with several attacks on AVs being recorded, including slashed tires, pelting with rocks, threats with firearms, and attempts to run AVs off the road [10].

In light of this persistent aversion, the current study investigates ways to improve peoples' perception of AVs and mitigate such objections to their presence on roads.

In general, people are averse to algorithms making decisions, showing a preference for human decision-makers even when their decisions are inferior to those produced by algorithms [11–13]. For example, Dietvorst and colleagues [11] found that people preferred forecasts produced by humans (as opposed to algorithms), even though humans made more mistakes and their forecasts were less accurate. Furthermore, Bigman and Gray [12] found that people are averse to algorithms making moral decisions (including in the context of AVs and driving), preferring that moral decisions be made by humans. One reason for peoples' aversion to AVs, and algorithmic decision-making in general, may be the "black box" nature of such algorithms. Many people may find it difficult to predict the decision-making processes of algorithms, leading to a lack of trust in these decisions even when they are accurate. Consistent with this explanation, Bigman and Gray [12] found that explicitly stating an algorithm's expertise attenuated the aversion.

Despite their promising safety record [2–5], the lack of transparency of AV algorithms represents a potential barrier to their eventual widespread adoption [14]. AV algorithms rely in part on machine learning, allowing these algorithms to be updated and improved independent of changes to their programming. Thus, the problem of algorithm transparency is unlikely to be fully solved as AV algorithms may come to lack transparency even for their programmers, let alone their passengers. Trust and predictability are major factors for increasing adoption of AVs [15]. Lacking knowledge into the decision-making processes of AVs, people may come to view AVs as unpredictable and untrustworthy and be averse to the use of AVs on roads. Furthermore, drivers with a distrust of AVs are more likely to behave unpredictably around them, making it more difficult for AV algorithms to determine and execute the safest course of action. Thus, enhancing public trust in AVs is important not only to allow for their eventual adoption but also to improve AV performance in situations where AVs and human drivers share the road.

Disagreements regarding how AVs should ethically behave represent another potential barrier to their widespread adoption [14]. That is, as AVs are introduced to roadways, they will face critical situations in which an AV algorithm will have to decide on actions that prioritize certain lives over others (e.g., a passenger versus a pedestrian). While many people may recognize a utilitarian approach (i.e., choosing the action that results in the least amount of harm) to be most moral, people purchasing AVs may nevertheless be uncomfortable with their vehicle not prioritizing their own safety [14]. Therefore, determining the most socially acceptable rules for AV action in such critical situations will also be crucial for their acceptance [16]. Nevertheless, regardless of the decision made, many people may simply be uncomfortable with AVs making moral decisions at all [12]. Thus, even consensus on *how* AVs should act across a host of critical situations does not guarantee that people will abandon their aversion to AVs making such decisions at all.

In a series of three studies, we used moral dilemmas to assess peoples' perceptions of both AVs and human drivers forced to make important moral decisions (i.e., do nothing and hit five pedestrians, or swerve into a single stranger). Across all studies, we investigate peoples' aversions to AVs, predicting that AVs will be judged more negatively (e.g., as less moral and more blameworthy) for performing the same action as a human driver. We also assess the perceived predictability of AVs compared to human drivers. While AVs dependence on algorithms is likely to make them behave *more* consistently than human drivers, the lack of transparency of these

algorithms may make them appear *less* predictable to human observers. In Studies 1 and 2, we examine whether highlighting either the consistency of AVs or the inconsistency of human drivers improves the perceived predictability of AVs and helps reduce AV aversion. In Study 2, we also assess whether AV aversion persists in situations where either an AV or human driver perform a skillful maneuver that helps bring about a positive outcome (i.e., zero people harmed). Finally, in Study 3, we assess how changes in pilot agency (i.e., an AV taking control of a vehicle from a human driver or vice versa) influence judgments of AVs and human drivers performing either deontological or utilitarian actions. Based on past work [17], we hypothesize that AVs would be judged especially harshly when taking control from a human driver.

METHODS

Participants

For all studies, participants were recruited from Amazon Mechanical Turk and received $0.50 upon completion of a 5-minute questionnaire. Participants were required to be residents of the United States and possess a Mechanical Turk HIT approval rating of 95% or greater to be eligible. Participation in one study excluded individuals from taking part in the other two studies.

We recruited initial samples of 1,407 (Study 1), 2,006 (Study 2), and 601 (Study 3) from Amazon Mechanical Turk. In each study, we excluded data from participants who failed to correctly respond to a comprehension check question. This criterion resulted in the exclusion of 268 participants from Study 1, 473 participants from Study 2, and 129 participants from Study 3. As such, Study 1 included a final sample of 1,139 participants (54.22% Female, M_{age} = 40.35, SD_{age} = 13.35), Study 2 a final sample of 1,533 participants (54.70% Female, M_{age} = 38.68, SD_{age} = 12.76), and Study 3 a final sample of 472 participants (50.53% Female, M_{age} = 40.00, SD_{age} = 12.90).

Materials and Measures

In all three studies, participants were presented with a vignette describing a scenario in which the pilot of a vehicle (either a human or AV) experiences a brake failure while driving down a hill. The pilot in this scenario was described as having a choice between keeping their vehicle on its present course and killing five pedestrians (Deontological Action) or altering their vehicle's path and killing a stranger getting out of their parked vehicle (Utilitarian Action; see Table 1). Following the presentation of this vignette, participants were told which choice the vehicle's pilot had made.

In Studies 1 and 2, one half of participants were presented with a second statement outlining either the consistency of AVs or the inconsistency of human drivers (determined by the pilot in the vignette; see Table 1, Studies 1 and 2). Study 2 also included a positive action condition, in which participants were told that either the human driver or the AV used precision steering to avoid killing the five pedestrians or the lone stranger (see Table 1, Study 2). Lastly, in Study 3, participants assigned to a Takeover condition were presented with a statement that described control of the vehicle being transferred from human driver to AV (or vice versa), followed by a description of the choice made by the pilot that assumed control (deontological or utilitarian; see Table 1, Study 3).

Scenario Description	
Base Vignette	*PILOT* is driving down a hill when the brakes on the vehicle fail. *PILOT's* vehicle will hit and kill five pedestrians in a crosswalk if the vehicle proceeds on its present course. In between *PILOT's* vehicle and the five pedestrians is a stranger getting out of their parked vehicle. The only way to save the lives of the five pedestrians is to hit the stranger's vehicle which will bring *PILOT's* vehicle to a stop. The stranger getting out of the parked vehicle will die if *PILOT* does this, but the five pedestrians will be saved.
Deontological Action	*PILOT* continues on the present course and hits the five pedestrians.
Utilitarian Action	*PILOT* hits the stranger and their parked vehicle.
Studies 1 and 2	
Human Driver Explanation	Human drivers can be inconsistent such that under the same conditions a human driver may not make the same decision every time. Based on his gut feelings, Michael… (*Utilitarian or Deontological Action*)
AV Explanation	Self-driving cars are consistent such that under the same conditions a self-driving car will make the same decision every time. Based on its algorithms, the self-driving vehicle… (*Utilitarian or Deontological Action*)
Study 2	
Human Driver Positive Action	Michael uses his fast reflexes and precision steering to avoid hitting the five pedestrians and the stranger getting out of their parked vehicle, instead hitting a decorative stone fountain causing only minimal property damage.
AV Positive Action	The self-driving vehicle uses its fast computing abilities and precision steering to avoid hitting the five pedestrians and the stranger getting out of their parked vehicle, instead hitting a decorative stone fountain causing only minimal property damage.
Study 3	
Human Driver Takeover	The self-driving feature on the car can be turned off, allowing the driver (Michael) to take control of the vehicle. Michael takes control of the vehicle… (*Utilitarian or Deontological Action*)
AV Takeover	Michael's vehicle is equipped with a self-driving feature that can automatically take control of the vehicle when it detects the possibility of a crash. The self-driving features takes control of Michael's vehicle… (*Utilitarian or Deontological Action*)

Table 1: Study Materials[1]

[1] The term *PILOT* was replaced with either the name "Michael' or "self-driving car" depending on whether participants were randomly assigned to the Human or AV Pilot condition. Materials under each study label are *in addition* to the base vignette.

Following the presentation of a vignette, participants were asked to judge the pilot who acted within the vignette on several dimensions. For each measure, participants made their judgments using a dichotomous 7-point scale with both endpoints labelled (i.e., Label 1/Label 2). In all studies, participants assessed the Goodness (Bad/Good), Morality (Immoral/Moral), Predictability (Unpredictable/Predictable), Harm (Caused No Harm/Caused a Great Deal of Harm), Blame (Deserves No Blame/Deserves a Great Deal of Blame), and Acceptability (Actions were Unacceptable/Actions were Acceptable) of a vehicle pilot. Additionally, Study 3 had participants judge the Trustworthiness (Untrustworthy/Trustworthy) of each described vehicle pilot. For all studies, participants' judgments of Goodness and Morality were averaged to create a Moral Perception composite[2] which was analyzed in place of these dimensions.

Design and Procedure

Study 1 used a 2 (Pilot: Human, AV) x 2 (Action: Deontological, Utilitarian) x 2 (Explanation: Explanation, No Explanation) between-subjects design in which all participants were presented with a single vignette featuring some combination of these three independent variables. Study 2 featured an almost identical design, however the addition of positive outcomes resulted in a 2 (Pilot: Human, AV) x 3 (Action: Deontological, Utilitarian, Positive) x 2 (Explanation: Explanation, No Explanation) between-subjects design. Lastly, Study 3 employed a 2 (Pilot: Human, AV) x 2 (Action: Deontological, Utilitarian) x 2 (Takeover: Takeover, No Takeover) mixed-factors design in which Pilot and Action were within-subject variables and Takeover a between-subjects variable. For all studies, participants were presented with either one (Studies 1 and 2) or two (Study 3) vignettes that described a scenario, and then judged the pilot acting within that scenario on a number of dimensions. Participants then completed a comprehension check question assessing if they understood key elements of the presented vignette(s), and concluded the questionnaire by answering two demographic questions (i.e., age and gender).

RESULTS

Data were analyzed using between-subjects (Studies 1 and 2) and mixed-factorial (Study 3) Analyses of Variance (ANOVA), with Type III sums of squares used to account for unequal cell sizes. Effect sizes were measured using partial eta squared (η_p^2) or Cohen's *d* when appropriate.

Study 1

Evidence of AV aversion was found in Study 1. Despite performing the same actions which resulted in the same outcomes, AVs were judged as less moral, less acceptable, as causing more harm, and as deserving of more blame compared to human drivers, $F(1,1131) > 3.50$, $p < .020$, $\eta_p^2 > .005$. We also observed main effects of Action, such that pilots described as performing the deontological action were judged as less moral, less acceptable, as causing more harm, and as more to blame, $F(1,1131) > 15.53$, $p < .001$, $\eta_p^2 > .014$.

[2] Judgments of Goodness and Morality were strongly correlated within all studies (Study 1: $r = .69$; Study 2: $r = .80$; Study 3: $r = .70$), justifying this composite.

We observed a Pilot by Explanation interaction for judgments of predictability, $F(1,1131) = 52.71$, $p < .001$, $\eta_p^2 = .045$ (see Figure 1A) and blame, $F(1,1131) = 7.37$, $p = .007$, $\eta_p^2 = .006$ (see Figure 1B). Follow-up independent-samples *t*-tests demonstrated that the inclusion of a statement highlighting the consistency of AVs resulted in AVs being judged as more predictable, $t(548) = 7.76$, $p < .001$, $d = 0.66$, and less blameworthy, $t(548) = 3.54$, $p < .001$, $d = 0.30$, compared to when no statement was provided. Conversely, highlighting the inconsistency of human drivers did not influence judgments of predictability or blame for human drivers (both *p's* > .068). No Pilot by Explanation interaction was observed for judgments of acceptability, moral perceptions, or harm (all *p's* > .553).

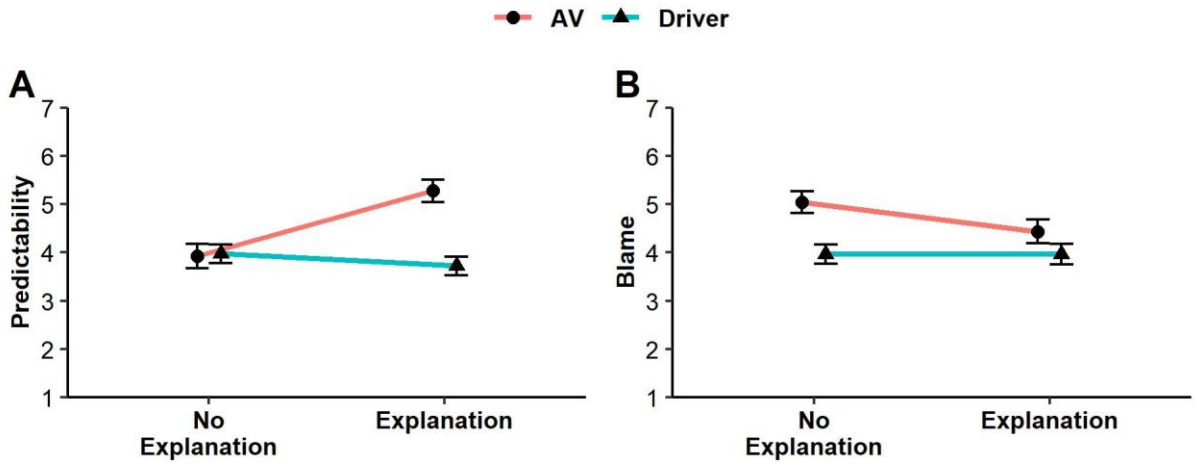

Figure 1 – Predictability (panel A) and Blame (panel B), as a function of Pilot and Explanation

Lastly, we examined the zero-order correlations between participants' judgments of predictability and other dependent variables within the AV Pilot condition. These analyses revealed positive associations between judgments of predictability and acceptability, $r(548) = .40$, $p < .001$, and moral perceptions, $r(548) = .45$, $p < .001$, along with negative associations between judgments of predictability and harm, $r(548) = -.14$, $p < .001$, and blame, $r(548) = -.30$, $p < .001$. Therefore, as participants judged AVs to be more predictable, they also tended to judge them as more acceptable, more moral, as causing less harm, and as less blameworthy. Notably, we observed the same associations when analyzing all judgments (i.e., including those from the Human Pilot condition).

Study 2

As in Study 1, participants' judgments revealed evidence of AV aversion, with AVs being judged as less moral, less acceptable, as causing more harm, and as being more blameworthy compared to human drivers, $F(1,1521) > 17.91$, $p < .001$, $\eta_p^2 > .012$. Study 2 also revealed main effects of Action, as positive actions were judged more favourably (i.e., as more moral, more acceptable, less harmful and less blameworthy) than deontological and utilitarian actions, $F(2,1521) > 570.0$,

$p < .001$, $\eta_p^2 > .248$. Consistent with Study 1, utilitarian actions were judged more favourably than deontological actions (all *p's* < .008).

Analyses also revealed a Pilot by Explanation interaction for judgments of predictability, $F(1,1521) = 90.05$, $p < .001$, $\eta_p^2 = .056$ (see Figure 2), and acceptability[3], $F(1,1521) = 3.94$, $p = .047$, $\eta_p^2 = .003$. Consistent with Study 1, follow-up independent-samples *t*-tests revealed that participants presented with an explanation highlighting the consistency of AVs judged AVs as more predictable than those not provided with this explanation, $t(753) = 10.10$, $p < .001$, $d = 0.74$. Similarly, participants presented an explanation highlighting the inconsistency of human drivers judged human drivers as less predictable than those not provided with this explanation, $t(776) = 2.88$, $p = .004$, $d = 0.21$. We did not observe a Pilot by Explanation interaction for judgments of blame, moral perceptions, or harm (all *p's* > .101).

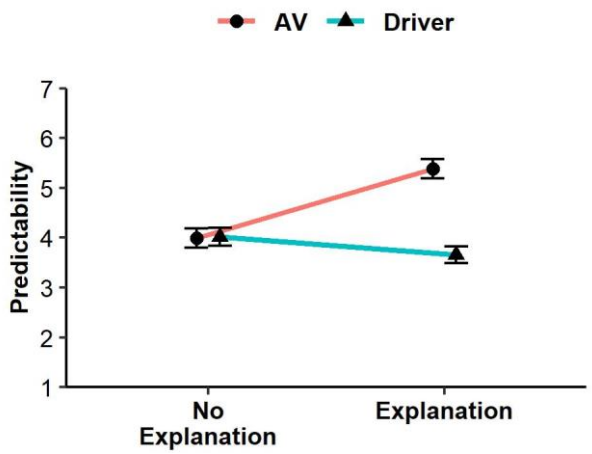

Figure 2 – Predictability, as a function of Pilot and Explanation

Additionally, Study 2 revealed a three-way interaction for judgments of acceptability, $F(2,1521) = 4.05$, $p = .018$, $\eta_p^2 = .005$ (See Figure 3), and moral perceptions, $F(2,1521) = 3.94$, $p = .020$, $\eta_p^2 = .005$ (see Figure 4). Decomposing these interactions, we found Pilot by Explanation interactions for deontological actions, $F(1,513) > 7.81$, $p < .005$, $\eta_p^2 > .015$, but not utilitarian, or positive actions (all *p's* > .171). That is, for deontological actions only, providing an explanation highlighting the consistency of AVs increased the acceptability, $t(259) = 2.37$, $p = .019$, $d = 0.29$, and moral perception, $t(259) = 2.24$, $p = .026$, $d = 0.28$, of AVs while explaining the inconsistency of human drivers did not influence the acceptability, $t(254) = 1.78$, $p = .076$, $d = 0.22$, or moral perception, $t(254) = 1.69$, $p = .091$, $d = 0.21$, of human drivers. Furthermore, participants' judgments of acceptability demonstrated less AV aversion (i.e., human drivers being judged as more acceptable than AVs) for positive actions, $F(1,499) = 5.31$, $p = .022$, $\eta_p^2 = .0.11$, compared to deontological and utilitarian actions, $F(1,513) = 43.78$, $p < .001$, $\eta_p^2 = .079$ and $F(1,509) = 20.12$, $p < .001$, $\eta_p^2 = .038$, respectively.

[3] We discuss this interaction in the following paragraph in the context of the three-way interaction observed for judgments of acceptability.

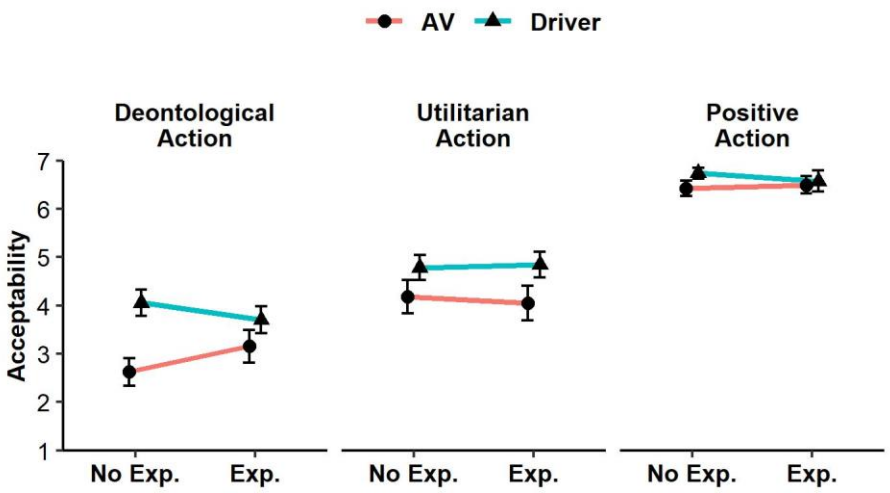

Figure 3 – Acceptability, as a function of Pilot, Action, and Explanation

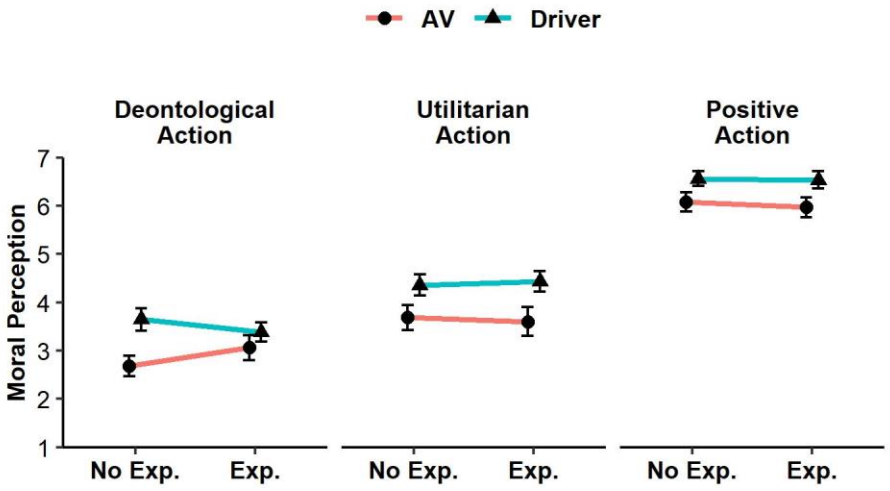

Figure 4 – Moral Perception, as a function of Pilot, Action, and Explanation

Finally, we conducted identical correlation analyses as described in Study 1. Consistent with Study 1, for AV Pilots we observed positive associations between judgments of predictability and acceptability, $r(753) = .25$, $p < .001$, and moral perceptions, $r(753) = .21$, $p < .001$. along with negative associations between judgments of predictability and blame, $r(753) = -.19$, $p < .001$. Thus, participants who perceived AVs as more predictable once again tended to judge AVs more favourably. Furthermore, as in Study 1, these associations were also observed when including judgments from the Human Pilot condition. However, unlike Study 1, we did not observe an

association between judgments of predictability and harm when analyzing either AV or all judgments (r > -.07, p > .05).

Study 3

We once again observed evidence of AV aversion in Study 3. Consistent with Studies 1 and 2, participants in Study 3 judged AVs as less moral, less trustworthy, less acceptable and more blameworthy compared to human drivers, $F(1,469) > 4.87$, $p < .028$, $\eta_p^2 > .010$. Furthermore, as in Studies 1 and 2, participants judged pilots performing the deontological action as less moral, less acceptable, less trustworthy, as causing more harm, and as more to blame compared to pilots performing the utilitarian action, $F(1,469) > 37.99$, $p < .001$, $\eta_p^2 > .074$.

Study 3 allowed us to assess the impact of takeover (i.e., control of the vehicle being transferred from human driver to AV or from AV to human driver) on participants' judgments. We observed a main effect of Takeover for moral perceptions and judgments of trustworthiness, $F(1,469) > 7.59$, $p < .007$, $\eta_p^2 > .015$. That is, pilots described as taking control of a vehicle immediately prior to acting within our scenario were judged as less moral and less trustworthy despite performing the same actions resulting in the same outcomes as pilots who were always in control of the vehicle. We also observed a Pilot by Takeover interaction for judgments of acceptability, $F(1,469) = 5.35$, $p = .021$, $\eta_p^2 = .011$ (see Figure 5). Participants judged the actions of human drivers taking control of an AV as less acceptable compared to human drivers performing the same actions while maintaining control of their vehicle (i.e., No Takeover condition), $t(469) = 3.02$, $p = .003$, $d = 0.28$. Conversely, the acceptability of AVs did not differ between Takeover and No Takeover conditions, $t(469) = 0.62$, $p = .530$, $d = 0.06$.

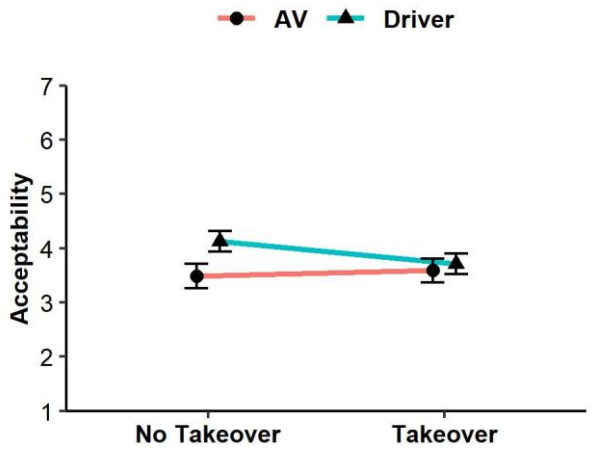

Figure 5 – Acceptability, as a function of Pilot and Takeover

Lastly, we examined the zero-order correlations between participants' judgments of predictability and our other dependent measures (i.e., acceptability, moral perceptions, trustworthiness, harm, and blame). Consistent with Studies 1 and 2, we observed positive associations between judgments of predictability and acceptability, $r(469) = .49$, $p < .001$, moral perceptions, $r(469) =$

.36, $p < .001$, and trustworthiness, $r(469) = .37$, $p < .001$, within the AV Pilot condition. Furthermore, we observed negative associations between judgments of predictability and harm, $r(469) = -.17$, $p < .001$, and blame, $r(469) = -.31$, $p < .001$. Thus, as in Studies 1 and 2, participants who perceived AVs as more predictable also tended to be less AV averse. Additionally, these associations were once again observed when analyzing all judgments, with the exception of the negative association between predictability and harm, which was reduced in magnitude and no longer statistically significant, $r(469) = -.08$, $p = .085$.

DISCUSSION & CONCLUSION

Consistent with past work demonstrating peoples' aversion to AVs and algorithmic decision-making [11–13, 15, 17–18], participants judged AVs as less acceptable, less moral, and more blameworthy compared to human drivers across our three studies. Furthermore, AVs were judged as less trustworthy than human drivers in Study 3 and as causing more harm in Studies 1 and 2. Notably, participants' negative judgments of AVs occurred despite AVs and human drivers being described as performing the same actions under the same circumstances. Thus, our results are consistent with prior work suggesting that AVs need to be considerably safer than human drivers before being granted the same level of acceptability [9].

Although a majority of the actions described in the current study were stated to result in a negative outcome (i.e., death), in Study 2 we assessed participants' judgments of AVs and human drivers in cases where the skilled maneuvering of either pilot was said to have avoided these outcomes. Past work has found evidence of algorithm aversion in situations where both the actions of humans and algorithms lead to positive outcomes [12]. Relatedly, people view the permissibility of algorithms making moral decisions as less than the permissibility of humans making these same decisions, even in situations where algorithms produce more favourable outcomes. Inconsistent with these findings, participants in Study 2 judged AVs performing positive actions more favourably than humans performing actions that lead to a negative outcome (i.e., deontological or utilitarian actions). Additionally, for judgments of acceptability, AV aversion was reduced when both AVs and human drivers performed the same positive action, as opposed to deontological or utilitarian actions. Thus, peoples' unwillingness to accept AVs may be especially pronounced when contemplating the actions of AVs and human drivers that lead to negative outcomes.

The lack of transparency of AV algorithms represents one potential explanation for AV aversion. This lack of transparency has been argued to be a potential barrier to the widespread adoption of AVs [14]. Because people don't understand the decision-making processes of AVs, they may view them as unpredictable and therefore untrustworthy. Given that trust and predictability have been discussed as major factors for increasing the adoption of AVs [15], perceptions of AVs as unpredictable and untrustworthy is likely to result in AV aversion and a hesitancy to purchase or share the roads with AVs.

In the current study, we assessed the perceived predictability of AVs and human drivers acting within a sacrificial moral dilemma. Across all three studies, participants judged AVs and human

drivers as similarly predictable[4]. Therefore, participants' aversion to AVs in the current study could not be explained by AVs appearing less predictable than human drivers. However, providing participants with a statement highlighting the consistency of AV algorithms (Studies 1 and 2) increased the predictability of AVs and, in some cases, reduced AV aversion. For example, in Study 1, explaining the consistency of AV algorithms increased the perceived predictability of AVs and reduced judgments of blame for AVs such that they were comparable to those given to human drivers. Furthermore, in Study 2, this explanation increased the perceived predictability of AVs as well as resulted in AVs being judged as more acceptable and moral when performing deontological actions, once again eliminating AV aversion in this context. Overall, highlighting the consistency of AVs was shown to increase the perceived predictability of AVs and showed some promise in reducing AV aversion by bringing peoples' judgments of AVs more in line with those given to human drivers. As such, future work should aim to further investigate how the impact of such messaging differs across a host of different contexts.

Recent work demonstrates the moral premium attached to predictable as opposed to unpredictable individuals, with people viewing predictable individuals as more moral and more trustworthy [19–20]. In line with such findings, we observed consistent associations between judgments of predictability and judgments of acceptability, morality, trustworthiness, blame, and harm. As participants came to perceive AVs as more predictable, they also came to view them as more acceptable, more moral, more trustworthy, less blameworthy, and as causing less harm. These associations ranged in magnitude ($.13 < r < .50$) yet were largely observed to be moderate in size. Furthermore, these relations largely remained when also analyzing judgments of human drivers, suggesting that as human drivers were perceived as more predictable, they too tended to be judged more favourably.

Finally, in Study 3, we examined participants' judgments of semi-autonomous vehicles in scenarios where an AV took control of a vehicle from a human driver (or vice versa) prior to performing either a utilitarian or deontological action that resulted in the deaths of 1 or 5 individuals. Previous research by McManus and Rutchick [17] found that AVs are judged especially harshly when taking control from human drivers, on account that this action reduces human agency. However, in Study 3, participants were found to be averse to takeover of any kind, regardless of which pilot was in control prior to the vehicles' brakes failing or while performing the utilitarian or deontological action. Moreover, participants judged the actions of a human driver as less acceptable after taking control from an AV, as compared to when no takeover occurred. The wholesale negative reaction to takeover actions in the current study represents a potential barrier to the trustworthiness, and eventual adoption of semi-autonomous vehicles. Nevertheless, future research assessing peoples' perceptions of takeover situations in other contexts is still needed. For example, it is unclear from the present study whether people judge takeover situations more negatively when the action performed leads to a positive outcome (e.g., the avoidance of an accident). Even so, preliminary results suggest that people are averse to situations in which agency over a situation is altered at key points of the decision-making process.

[4] That is, when not explicitly presented a statement describing the consistency of AV algorithms or the inconsistency of human drivers.

The wide-spread adoption of AVs has the potential to reduce the frequency and severity of motor vehicle accidents [2–3]. However, peoples' aversion to AVs (e.g., lack of trust) represents an important barrier to their implementation and ability to improve roadway safety. Across three studies, participants displayed an aversion to AVs such that they judged AVs more negatively compared to human drivers performing the same actions under the same circumstances. We hypothesized that one reason for this aversion was the lack of transparency of AV algorithms, resulting in AVs appearing unpredictable. As such, we examined whether highlighting the consistency of AV algorithms could enhance perceptions of AV predictability and reduce AV aversion. Such messaging did improve perceptions of AV predictability, and we find some evidence regarding its ability to reduce AV aversion.

However, the consistency of algorithms may not be the only determinant in how AVs are perceived. For instance, future AV technologies will likely rely on cameras—similar to those used currently to detect lane departures and forward collisions—to navigate the environment autonomously. As such, important elements within the environment (e.g., lane markings, signage) may come to impact AV action and performance. Yet troublingly, lane markings, signage, and even traffic lights still vary between jurisdictions, and in colder climates these elements can even become obstructed by snow, ice, and salt [21]. Furthermore, efforts to upgrade and standardize such features so that they are universally interpretable by AV technologies will likely occur more slowly than the rapid advancement of AV technology itself. Consequently, the perceived predictability of AVs could be reduced, and AV aversion increased, by AVs reliance on seemingly inconsistent and unreliable environmental features. Thus, future research investigating the efficacy of messaging should not only aim to investigate the perceived predictability and acceptability of AVs by explicitly contrasting their superior capabilities and consistency against the fallibility (and inconsistency) of human drivers (e.g., their tendency to become distracted, mind-wander, etc. [22–23]), but also evaluate opinion when failure to execute a favourable outcome rests solely on limitations within the driving environment itself.

REFERENCES

[1] NHTSA, Automated Vehicles for Safety, *National Highway Traffic Safety Administration,* 2020.

[2] GAO, P., HENSLEY, R., ZIELKE, A., A Road Map to the Future for the Auto Industry, *McKinsey Quarterly,* 1-11, October 2014.

[3] BELLIS E., PAGE, J., National Motor Vehicle Crash Causation Survey (NMVCCS), *SAS Analytical Users Manual,* No. HS-811 053, 232 p., 2008.

[4] BLANCO, M., ATWOOD, J., RUSSELL, S., TRIMBLE, T., MCLAFFERTY, J., PEREZ, M., Automated Vehicle Crash Rate Comparison Using Naturalistic Data, *Virginia Tech Transportation Institute,* 77 p., 2016.

[5] TEOH, E.R., KIDD, D.G., Rage Against the Machine? Google's Self-Driving Versus Human Drivers, *Journal of Safety Research,* 63, 57-60, 2017.

[6] DAVIES, A., Tesla's Latest Autopilot Death Looks Just Like a Prior Crash, *WIRED*, May 2019.

[7] EDMONDS, E., Three in Four Americans Remain Afraid of Fully Self-Driving Vehicles, *AAA NewsRoom,* March 2019.

[8] SAID, C., Video Shows Uber Robot Car in Fatal Accident Did Not Try to Avoid Woman. *SFGATE,* March 2018.

[9] LIU, P., YANG, R., XU, Z., How Safe is Safe Enough for Self-Driving Vehicles?, *Risk Analysis,* 39, 2, 315-325, 2019.

[10] ROMERO, S., Wielding Rocks and Knives, Arizonans Attack Self-Driving Cars, *The New York Times,* December 2018.

[11] DIETVORST, B.J., SIMMONS, J.P., MASSEY, C., Algorithm Aversion, People Erroneously Avoid Algorithms After Seeing Them Err, *Journal of Experimental Psychology: General,* 144, 1, 114-126, 2015.

[12] BIGMAN, Y.E., GRAY, K., People Are Averse to Machines Making Moral Decisions, *Cognition,* 181, 21-34, 2018.

[13] NISZCZOTA, P., KASZÁS, D., Robo-Investment Aversion, *PLoS ONE,* 15, 9, 1-19, 2020.

[14] SHARIFF, A., BONNEFON, J.F., RAHWAN, I., Psychological Roadblocks to the Adoption of Self-Driving Vehicles, *Nature Human Behaviour*, 1, 10, 694-696, 2017.

[15] CHOI, J.K., JI, Y.G., Investigating the Importance of Trust on Adopting an Autonomous Vehicle, *International Journal of Human-Computer Interaction,* 31, 10, 692-702, 2015.

[16] BONNEFON, J.F., SHARIFF, A., RAHWAN, I., The Social Dilemma of Autonomous Vehicles, *Science*, 352, 6293, 1573-1576, 2016.

[17] MCMANUS, R.M., RUTCHICK, A.M., Autonomous Vehicles and the Attribution of Moral Responsibility, *Social Psychological and Personality Science,* 10, 3, 345-352, 2019.

[18] KOO, J., KWAC, J., JU, W., STEINERT, M., LEIFER, L., NASS, C., Why Did My Car Just Do That? Explaining Semi-Autonomous Driving Actions to Improve Driver Understanding, Trust, and Performance, *International Journal on Interactive Design and Manufacturing*, 9, 4, 269-275, 2015.

[19] WALKER, A.C., TURPIN, M.H., FUGELSANG, J.A., BIALEK, M., Better the Two Devils You Know, Than the One You Don't: Predictability Influences Moral Judgments of Immoral Actors, *Journal of Experimental Psychology,* 97, 104220, 2021.

[20] TURPIN, M.H., WALKER, A.C., FUGELSANG, J.A., SOROKOWSKI, P., GROSSMAN, I., BIALEK, M., The Search for Predictable Moral Partners: Predictability and Moral (Character) Preferences, *Journal of Experimental Psychology*, 97, 104196, 2021.

[21] CARLSON, P., Traffic Control Devices: Considerations to Support Automated Vehicle Deployment, *Transport Canada*, 29 p., 2021

[22] WALKER, H.E.K., TRICK, L.M., Mind-Wandering While Driving: The Impact of Fatigue, Task Length, and Sustained Attention Abilities, *Transportation Research Part F: Traffic Psychology and Behaviour,* 59, 81-97, 2018.

[23] WALKER, H.E.K., ENG, R.A., TRICK, L.M., Dual-Task Decrements in Driving Performance: The Impact of Task Type, Working Memory, and the Frequency of Task Performance, *Transportation Research Part F: Traffic Psychology and Behaviour,* 79, 185-204, 2021.