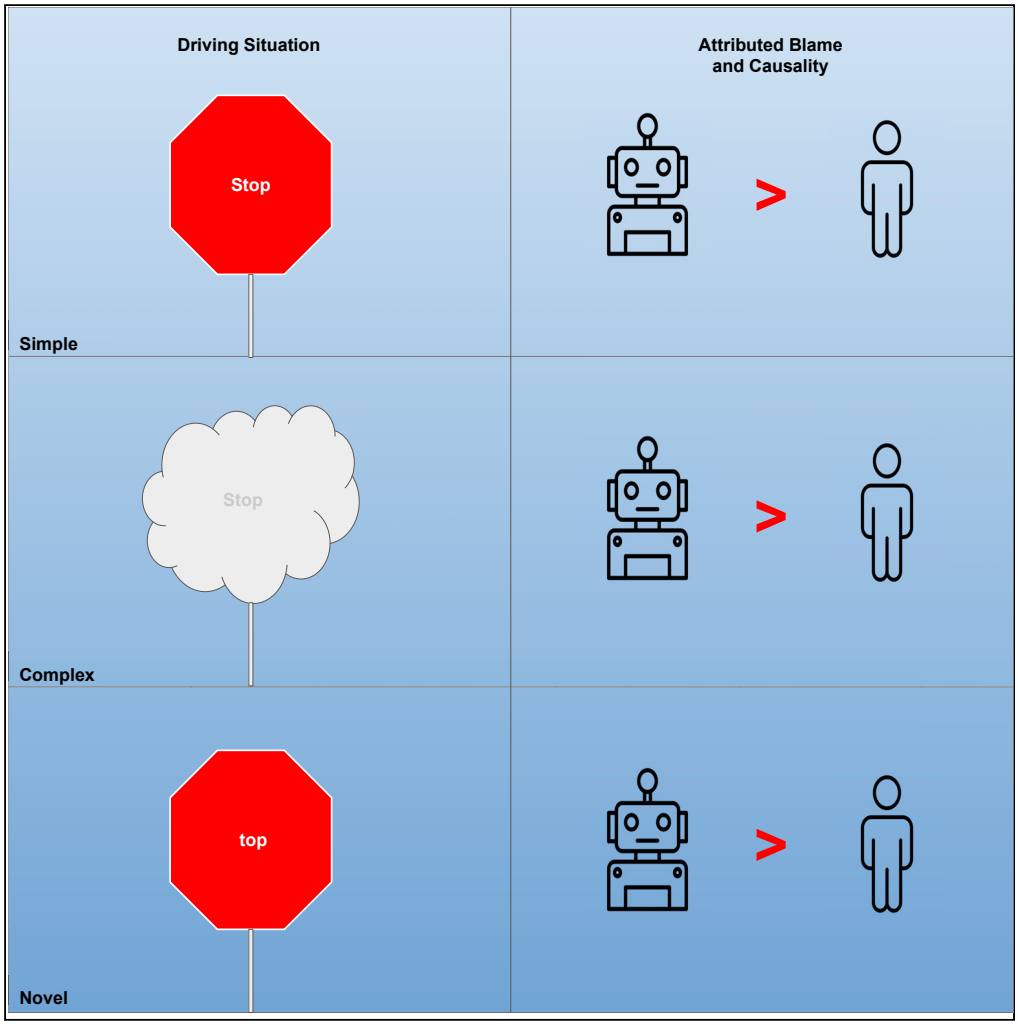
iScience

CellPress
OPEN ACCESS

Article

Blaming automated vehicles in difficult situations



Matija Franklin, Edmond Awad, David Lagnado

matija.franklin@ucl.ac.uk

HIGHLIGHTS

Attributed blame to machine and human drivers is sensitive to situation difficulty

Mistakes in simple situations receive more blame than in novel or complex situations

Machine drivers receive more blame, across different situations

Franklin et al., iScience 24, 102252
April 23, 2021
https://doi.org/10.1016/j.isci.2021.102252

iScience

CellPress
OPEN ACCESS

Article

Blaming automated vehicles in difficult situations

Matija Franklin,[1,3,*] Edmond Awad,[2] and David Lagnado[1]

SUMMARY

Automated vehicles (AVs) have made huge strides toward large-scale deployment. Despite this progress, AVs continue to make mistakes, some resulting in death. Although some mistakes are avoidable, others are hard to avoid even by highly skilled drivers. As these mistakes continue to shape attitudes toward AVs, we need to understand whether people differentiate between them. We ask the following two questions. When an AV makes a mistake, does the perceived difficulty or novelty of the situation predict blame attributed to it? How does that blame attribution compare to a human driving a car? Through two studies, we find that the amount of blame people attribute to AVs and human drivers is sensitive to situation difficulty. However, while some situations could be more difficult for AVs and others for human drivers, people blamed AVs more, regardless. Our results provide novel insights in understanding psychological barriers influencing the public's view of AVs.

INTRODUCTION

Once properly prepared and finalized to deploy on the roads, automated vehicles (AVs) are expected to bring many benefits, such as decreasing the rate of car crashes (Gao et al., 2014), reducing pollution (Spieser et al., 2014), and increasing traffic efficiency (van Arem et al., 2006). Assuming that AVs will overcome all remaining technical challenges before they are ready to deliver these benefits, while exhibiting no serious drawbacks, their deployment on a larger scale would be beneficial. However, these benefits will not be realized if people are not ready to buy them, and various considerations contribute to the public's aversion to adopting this technology.

Understanding people's attitudes is key to identifying these considerations, and working to address any potential concerns (Shariff et al., 2017; Schlögl et al., 2019; Sun and Medaglia, 2019; Bonnefon et al., 2020; Dellaert et al., 2020). The public's views and trust toward AVs is a major factor that predicts adoption of autonomous vehicles (Lee and Moray, 1992, 1994; Gefen et al., 2003; Carter and Bélanger, 2005). Evidence suggests that people require AVs to be multiple orders of magnitude safer than human drivers (Liu et al., 2019). As argued in (Awad et al., 2020), negative public reaction may result in inflated prices of this technology (Geistfeld, 2017) and may shape how a tort-based regulatory scheme would turn out, both of which can influence the rate of adoption.

In such cases of high stakes (safety of life), human attitude is mainly shaped by situations of failure. An autonomous vehicle may navigate its way successfully on the roads for long periods of time but will still be slammed for failing to avoid a crash in one situation. This asymmetric effect of performance on the public's attitude is amplified by the wide coverage of the few crashes by AVs, compared to the coverage of successful performance or achieved milestones by these AVs, and also compared to crashes by human drivers (Lambert, 2018). The strong reactions these few crashes have elicited point to the importance of focusing on mistakes and the failure situations to understand the publics' attitude.

Understanding how we react to mistakes by machines (as compared to those by humans) is not an easy task. There is strong evidence that people react differently to mistakes made by machines and humans (Dietvorst et al., 2014, 2015; Malle et al., 2015; Awad et al., 2020). There are also reasons to believe that people assign blame differently based on the difficulty of encountered situations. Complicating matters, perceived difficulty of the situation may vary depending on the agent behind the steering wheels. For example, a drunk person jumping in front of a car may be considered a difficult situation for a human driver but not for a machine driver (Goodall, 2016). Likewise, a novel situation in which the only way to overtake a stationary

[1]Department of Experimental Psychology, University College London, London WC1E 6BT, UK

[2]Department of Economics, University of Exeter Business School, Exeter EX4 4PU, UK

[3]Lead contact

*Correspondence: matija.franklin@ucl.ac.uk

https://doi.org/10.1016/j.isci.2021.102252

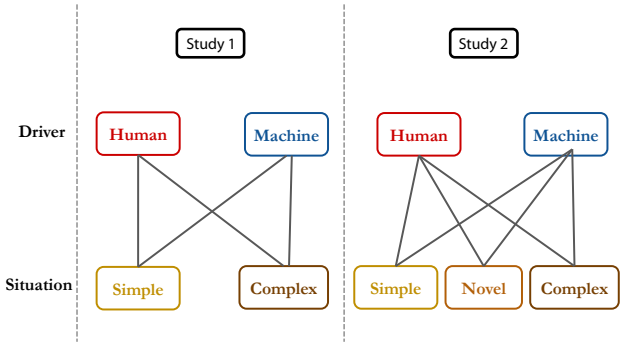


Figure 1. Experimental design of the two studies

The edges connecting *Driver* and *Mistake* represent the experimental groups participants were allocated to, with a total of four experimental groups for study 1: human driver in a simple situation, human driver in a complex situation, machine driver in a simple situation and machine driver in a complex situation; and six experimental groups for study 2 – same as in study 1 with the addition of two new experimental groups: human driver in a novel situation and machine driver in a novel situation.

vehicle is to illegally cross the central line may be deemed more difficult for a machine driver than for a human driver.

In this study, we focus on two questions: (1) when an automated car makes a mistake, does the perceived difficulty or novelty of the situation predict blame attributed to it? (2) How does that blame attribution compare to a human driving a regular car?

To answer these questions, we devise two studies that look at mistakes made by human drivers and machine drivers in driving situations that span different levels of difficulty (see Figure 1). Specifically, we consider three types of situations: (a) simple driving situations: those that most humans would consider easy to navigate without making mistakes (most of the time). (b) complex situations: those that add extra layers of difficulty or complication to the simple driving situation, requiring a higher level of competence to navigate while avoiding making mistakes, and (c) novel situations: those that are less likely to be encountered while driving (than simple or complex situations) and require novel inferences and actions that would not have been part of pre-training. The consideration of complex and novel situations here is crucial, given that what is deemed as difficult for a human driver (e.g., split-second decisions) can be an easy task for an AV. On the other hand, a novel situation (e.g., having to make an illegal move to overtake a stationary vehicle) could prove more challenging for an AV than for a human driver (Marcus and Davis, 2019).

RESULTS

In study 1, 198 participants were allocated randomly to one of four conditions: (driver: human vs. machines): x (situation: simple vs. complex). Each participant read six different stories of a mistake by a driver resulting in a crash, and then assigned scores of blame and causality to the driver. These scores were summed up into two separate blame and causality scores. The descriptive statistics and the individuals' blame scores are available in Figure 2 (causality scores are very similar).

Our results show that machines receive more blame and causality attribution than humans for either type of mistakes and that humans and machines are blamed more for mistakes in simple situations (see Figure 2). For blame attributions, results from a 2 × 2 ANOVA (driver x situation) show that machine drivers are blamed significantly more than human drivers for doing the same mistakes [$F(196, 2) = 5.82$, $p = .011$] and that all drivers get more blame for mistakes in simple, rather than complex, situations [$F(196, 2) = 6.73$, $p = .004$]. These results were replicated for causality attributions with machines receiving higher causality attributions [$F(196, 2) = 6.18$, $p = .008$] and drivers committing mistakes in simple situations being perceived as more causally responsible [$F(196, 2) = 4.66$, $p = .045$].

In study 2, 317 participants were allocated randomly to one of six conditions: (driver: human vs. machines): x (situation: simple vs. novel vs. complex). Each participant read five different stories of a mistake by a driver resulting in a crash, and then assigned scores of blame and causality to the driver. Participants also rated

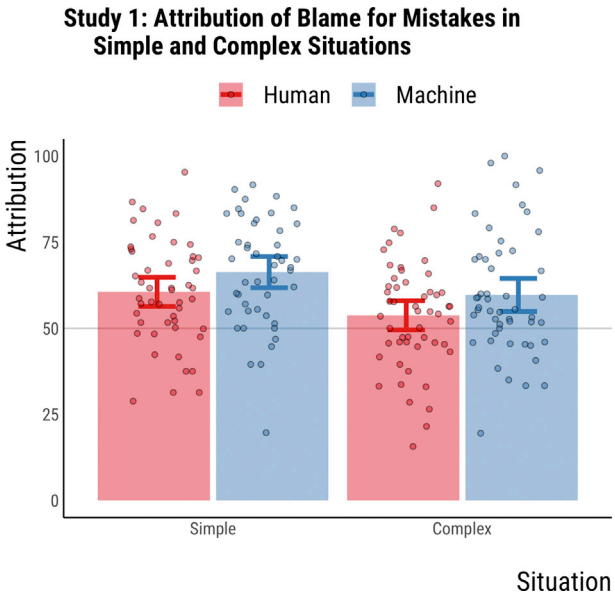


Figure 2. Attribution of blame for mistakes in simple and complex situations in study 1

Data from study 1 (n = 198). Participants were randomly allocated to one of four groups. The x axis represents the situation difficulty (simple vs. complex). The y axis represents blame attribution. Blue and red bars represent the mean blame attribution to the machine driver (AV) and the human driver, respectively. Error bars represent 95% confidence intervals of the mean. Each circle represents an individual's blame score (averaged over six stories). Machine drivers are blamed more than human drivers in total and across the two types of scenarios [$F(196, 2) = 6.17$, $p = .014$]. Machine and human drivers are blamed for making mistakes more in simple situations than in complex situations [$F(196, 2) = 8.36$, $p = .004$]. Data are represented as mean ± SEM.

the driving situation in terms of novelty and difficulty. Blame judgments and causality attributions were summed up into two separate blame and causality scores. The descriptive statistics for blame scores are available in Figure 3 (causality scores are very similar).

Our results show that machines receive more blame and causality attribution than humans for mistakes in novel and complex mistakes (see Figure 3). Furthermore, humans and machines are blamed the most for mistakes in simple situations, followed by novel and then complex situations. For blame attributions, results from a 2 × 3 ANOVA (driver x situation) show that overall machine drivers are blamed significantly more than human drivers for doing the same mistake [$F(314,3) = 7.81$, $p = .006$] and that all drivers receive significantly different levels of blame for mistakes in different driving situations [$F(314,3) = 60.75$, $p < .001$]. Post hoc comparisons using the Tukey honestly significant difference (HSD) test indicated that the mean blame score in simple situations ($M = 367$, $SD = 75.98$) was significantly higher than in novel ($M = 305.74$, $SD = 101.98$) and complex ($M = 234.24$, $SD = 86.5$) situations. Further, the mean blame score in novel situations was significantly higher than in complex situations. The patterns in this ANOVA were replicated in ANOVAs performed for individual items (See S2).

These results were replicated for causality attributions with machines receiving higher causality attributions [$F(314,3) = 11.66$, $p < .001$] and drivers rated as differentially causal for different scenarios [$F(314,3) = 67.48$, $p < .001$]. Post hoc comparisons using the Tukey HSD test indicated that the mean causality score in simple situations ($M = 383.3$, $SD = 71.2$) was significantly higher than in novel ($M = 313.1$, $SD = 99.65$) and complex ($M = 251.38$, $SD = 79.87$) situations. Further, the mean causality score in novel situations was significantly higher than in complex situations. The patterns in this ANOVA were replicated in ANOVAs performed for individual items (See S2).

To test for participants' perception of how difficult human and machine drivers would find particular situations, and to examine the success of the experimental manipulation of the study's items, we observed people's ratings of situations' difficulty and novelty. Specifically, the five separate ratings that were given by participants for each item was summed into a novelty and difficulty score. The descriptive statistics (see Figure 4) imply that the items elicited the desired response, with items describing novel situations

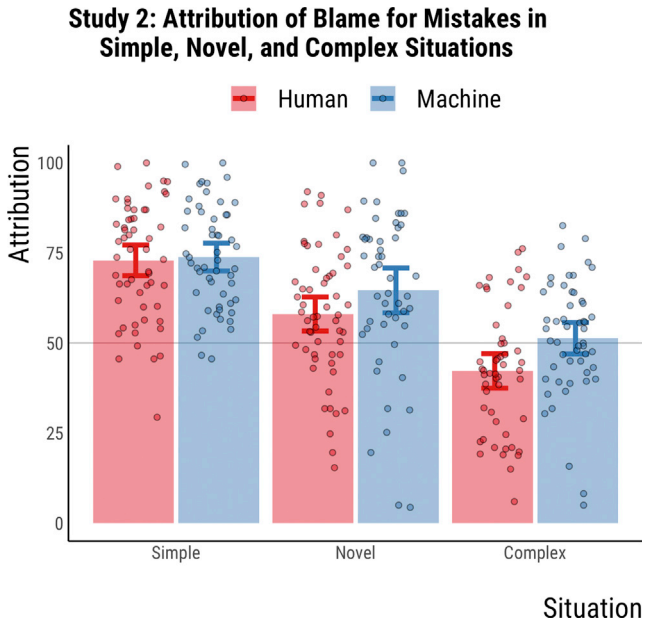


Figure 3. Attribution of blame for mistakes in simple, novel, and complex situations in study 2

Data from study 2 (n = 317). Participants were randomly allocated to one of six groups. The x axis represents the situation difficulty (simple vs. novel vs. complex). The y axis represents blame attribution. Blue and red bars represent the mean blame attribution to the machine driver (AV) and the human driver, respectively. Error bars represent 95% confidence intervals of the mean. Each circle represents an individual's blame score (averaged over five stories). Machine drivers are blamed more than human drivers in total and across two types of scenarios (inconclusive for simple situations) [$F(315, 2) = 4.99$, $p = .026$]. There were significant differences in blame across driving situations [$F(314, 3) = 59.05$, $p < .001$], with mean score for blame in simple situations being significantly higher than mean scores for blame in novel [differences in means = 61.26, 95% CI: [32.69, 89.84], $p < .001$] and complex situations [differences in means = 132.76, 95% CI: [103.98, 161.54], $p < .001$]. Mean scores for blame in novel situations were higher than those in complex situations [differences in means = 71.49, 95% CI: [42.58, 100.41], $p < .001$]. Data are represented as mean ± SEM.

having the highest novelty scores, and items describing complex situations having the highest difficulty scores.

For novelty scores, results from a 2 × 3 ANOVA (driver x situation) show that identical driving situations are scored as more novel if they have a machine driver in them rather than a human driver [$F(314,3) = 9.65$, $p = .002$] and that different driving situations produce a significant difference in novelty scores [$F(314,3) = 4.15$, $p = .017$]. Post hoc comparisons using the Tukey HSD test indicated that the mean novelty score in novel situations ($M = 248.68$, $SD = 105.35$) was significantly higher than in simple situations ($M = 213.62$, $SD = 108.63$).

For difficulty scores, results from a 2 × 3 ANOVA (driver x situation) show that identical driving situations are scored as more difficult if they have a machine driver in them rather than a human driver [$F(314,3) = 14.98$, $p < .001$] and that different driving situations produce a significant difference in difficulty scores [$F(314,3) = 20.99$, $p < .001$]. Post hoc comparisons using the Tukey HSD test indicated that the mean difficulty score in complex situations ($M = 275.3$, $SD = 83.05$) was significantly higher than in simple ($M = 188.93$, $SD = 110.95$) and novel situations ($M = 236.17$, $SD = 100.17$). Further, the mean difficulty score in novel situations was significantly higher than in simple situations.

Further, we examined whether trust in other drivers or AVs were predictive of blame judgments and causal attributions. For this, four separate linear regressions were conducted – two for participants in machine driving scenarios and another two for participants in human driving scenarios. For groups with human drivers, trust in other drivers did not significantly predict blame judgments and causal attributions. For groups with machine drivers, the results of two linear regressions showed that trust in AVs significantly predicted blame judgment [$F(153, 2) = 4.06$, $p = .046$, $R^2 = .026$] and causal attribution [$F(153, 2) = 5.19$, $p = .024$, $R^2 = .033$]. The results show that people's trust in AVs predicts their judgments and attributions of AVs.

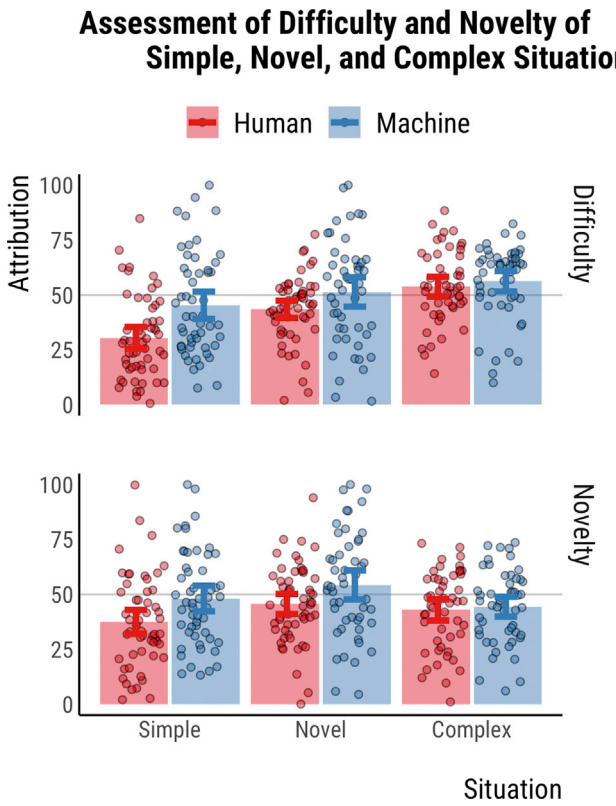


Figure 4. Assessment of difficulty and novelty of simple, novel, and complex situations in study 2

Data from study 2 (n = 317). Participants were randomly allocated to one of six groups. The x axis represents the situation difficulty (simple vs. novel vs. complex).The y axis represents difficulty or novelty assessment. Blue and red bars represent the mean assessment to the described scenarios featuring a machine driver (AV) and a human driver, respectively. Error bars represent 95% confidence intervals of the mean. Each circle represents an individual's difficulty or novelty scores (averaged over five stories). For situations featuring a human driver, the mean difficulty score in complex situations is significantly higher than in novel [differences in means = −51.46, 95% CI: [-89.89, −13.03], p = .005] and simple [differences in means = −116.65, 95% CI: [-155.24, −78.05], p < .001] situations. For situations featuring a machine driver, the mean difficulty score in complex situations is significantly higher than in simple situations [differences in means = −54.61, 95% CI: [-104.34, −4.88], p = .028], and the mean novelty score in novel situations is significantly higher than in complex situations [differences in means = 49.75, 95% CI: [.92, 98.58], p = .045]. Data are represented as mean ± SEM.

Finally, we examined participants' expectations of whether they thought a mistake was going to happen or not after completing the items (human: M = 5.02, SD = 1.45; machine: M = 5.35, SD = 1.25). The results of an independent sample T test show that people's expectations for a mistake were significantly larger for machines than humans [d(315,2) = -.24, p = .017]. Expectations were then used to predict participants' judgments and attributions, using two separate linear regressions. When using blame attribution as the dependent variable, the linear regression produced a significant effect [F(315,2) = 4.63, p = .032, R^2 = .014]. This finding was replicated when causality attribution was the dependent variable [F(315,2) = 9.09, p = .003, R^2 = .028]. The results show that people's expectations predict their judgments and attributions.

We reviewed participants' qualitative responses to how they made their choices. The three main themes were that participants' own experience with driving, the situation they were analyzing, and that the driver they were judging informed their decisions.

DISCUSSION

Before being deployed on the road, AVs need to reach a satisfactory level of competence. What is satisfactory, however, may be measured against satisfactory levels of human driving. But comparisons between humans and machines are hard to make, since each is challenged differently, even when facing the same set

of situations. While AVs operating based on state-of-the-art technology can navigate fast-calculation, quick-reaction types of situations better than humans, they still struggle with novel (yet seemingly simple from a human perspective) situations. Because of this, one would expect that people will blame machine drivers less than human drivers for making mistakes in novel situations but blame them more in complex situations that rely on handling more information or calculation.

Surprisingly, we find that this is not the case. In two studies, we find that participants blamed machine drivers more than human drivers for mistakes in both complex and novel situations. The differences in blame toward machines and humans may seem small on a scale 1-100 (meaning only few more people would find machines more blameworthy than humans, than vice versa). However, such differences are likely to map to practical turning points in real life. Our studies are done with a group of independent individuals faced with neutral description of scenarios. This ignores two factors: (1) media influence and (2) social influence, both of which are expected to magnify the difference. As for (1), If our sample is any representation of journalists, this will be reflected in more blame-the-machine biased articles, that are read by many more individuals. As for (2), it is plausible to assume that social influence of judgment happens according to a majority-voting model (majority of your neighbors determine your "state" with high probability p_{agree}) (Campos et al., 2003; Liggett, 2012), and that influence propagates over our social networks, represented as "small-world" networks (Watts and Strogatz, 1998; Amaral et al., 2000). In such settings, "who is blamed more?" matters more than "how much blamed more?" in shaping the final collective judgment (Ray et al., 2021).

There are multiple possible explanations that may help understand these findings. The first possibility is that people were not sensitive to the difficulty or novelty of these situations in a way that follows the desired experimental manipulations. However, this is unlikely, given that (as illustrated in Figure 4) participants assigned higher novelty scores for novel situations than for simple or complex situations when faced by machines. They also assigned higher difficulty scores for complex situations than the other two when faced by human drivers. Finally, novel and complex situations were considered more difficult than simple situations when faced by machine drivers.

One factor that can influence blame attribution is whether responsibility is shared. Previous work showed that people factor in the role of other agents in the broader system when making judgments (Lagnado et al., 2013). Given this, if a machine and a human perform an identical action with the same consequence, people might view them differently in terms of their causality and blameworthiness due to the other agents that are somewhat responsible for their behavior. However, this also seems like an unlikely explanation for our findings. Compared to humans, one would expect that machines should be "sharing" the blame with more agents, such as developers, designers, data scientists, data sets and manufacturers, a problem that has been identified before as "AI Responsibility Gap" (Matthias, 2004), and "moral crumple zone" (Elish, 2019).

One possible explanation is that participants perceive machines to be more competent at driving than humans. Expectations of someone's skills influence people's blame judgment for an outcome (Gerstenberg et al., n.d.). Specifically, when one has a high prior expectation of how someone will behave, they will see them as more blameworthy if they underperform and cause a negative outcome. However, this explanation is refuted by our post-assessment questions which found participants expressing higher likelihood of making a mistake (in the considered situations) for machine drivers than for human drivers. This runs counter to the literature on the public's risk acceptance of AVs, which shows that the public expects AVs to be significantly safer than human drivers, feel reasonably safe riding in an AV, and would allow AVs on public roads (Nees, 2019). Our participants' expectation of failure can be seen as an expectation of a driver's ability to avoid a mistake i.e., higher perceived competence at driving for humans.

Another possible explanation is that participants hold machines to higher standards than humans for the task of driving. In this case, even when people perceive machines as less competent drivers than humans, they still want them to follow higher standards before they are allowed to drive us. Our finding that for AVs, higher trust predicts higher blame may somewhat support this conclusion. Given that we do not have data that refutes or confirms this possibility, this remains as a possible explanation for why machines receive more blame than humans.

Finally, the answer may lie in the attribution of causal responsibility. People judge agents as more blameworthy for an outcome if they see them as more causally responsible for that outcome (Gerstenberg and Lagnado, 2010). In this case, if machines are seen as more causally responsible for the described mistakes, we would expect them to blame them more for these mistakes. Our data show that causal attribution results mirror the blame attributions for all conditions. However, this does not provide a satisfactory explanation. It only shifts the focus to causal responsibility: Why do people find machines more causally responsible than humans irrespective of the difficulty of the situation?

The immediate future is likely to see machines assuming new instrumental roles in industry and governance. This has, and will, lead to new situations where the developers of such intelligent machines are not able to fully predict their machines behavior and thus mistakes. The current study contributes to the existing literature (Malle et al., 2015; Awad et al., 2020; Bennett et al., 2020; Hidalgo et al., 2021) seeking to understand how situations of failure shape the public's attitude toward machines. Exploring how this is likely to unfold is a crucial step forward toward realizing the potential benefit of this technology.

Limitations of the study

The study has three limitations pertaining to its participants, measures and explanations it draws from its results. First, although our sample size met the requirements of a power analysis (see Participants section in Methods), a larger sample size across the two studies would have made for more conclusive results. Further, as our sample was recruited from MTurk and included participants from the UK, it was not fully representative. Second, the measure used for people's expectations of whether a mistake was about to happen was ad hoc, thus participants' expectations could have been influenced by the experimental items which described the driving situations. Finally, the current study cannot fully provide a definitive explanation for why machines are blamed more across all driving situations.

Resource availability

Lead contact

Further information and requests for experimental materials and data should be directed to and will be fulfilled by the lead contact, Matija Franklin (matija.franklin@ucl.ac.uk).

Materials availability

All items used in the online experiment are available from the lead contact without restriction.

Data and code availability

All data generated or analyzed during this study are currently available in the Figshare repository. The data from study 1 are available here https://figshare.com/articles/dataset/Blaming_Automated_Vehicles_Study_1_/12982085. The data from study 2 are available here https://figshare.com/articles/dataset/Blaming_Automated_Vehicles_Study_2_/12982103.

METHODS

All methods can be found in the accompanying Transparent methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.102252.

ACKNOWLEDGMENTS

The authors would like to thank members of the Causal Cognition Laboratory, UCL for their support. The study was funded by University College London.

AUTHOR CONTRIBUTIONS

M.F., E.A., and D.L. designed the experiment and wrote the paper. M.F. launched the experiment and analyzed the results. E.A. created all of the presented figures.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 21, 2020
Revised: January 15, 2021
Accepted: February 24, 2021
Published: April 23, 2021

REFERENCES

Amaral, L.A., Scala, A., Barthelemy, M., and Stanley, H.E. (2000). Classes of small-world networks. Proc. Natl. Acad. Sci. U S A *97*, 11149–11152.

van Arem, B., van Driel, C.J.G., and Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans. Intell. Transport. Syst. *7*, 429–436, https://doi.org/10.1109/tits.2006.884615.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J.B., Shariff, A., Bonnefon, J.F., and Rahwan, I. (2020). Drivers are blamed more than their automated cars when both make mistakes. Nat. Hum. Behav. *4*, 134–143.

Bennett, J.M., Challinor, K.L., Modesto, O., and Prabhakharan, P. (2020). Attribution of blame of crash causation across varying levels of vehicle automation. Saf. Sci. *132*, 104968, https://doi.org/10.1016/j.ssci.2020.104968.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2020). The moral psychology of AI and the ethical Opt-out problem. Ethics Artif. Intelligence, 109–126, https://doi.org/10.1093/oso/9780190905033.003.0004.

Campos, P.R.A., de Oliveira, V.M., and Moreira, F.G.B. (2003). Small-world effects in the majority-vote model. Phys. Rev. E Stat. Nonlin Soft Matter Phys. *67*, 026104.

Carter, L., and Bélanger, F. (2005). The utilization of e-government services: citizen trust, innovation and acceptance factors. Inf. Syst. J. *15*, 5–25, https://doi.org/10.1111/j.1365-2575.2005.00183.x.

Dellaert, B.G.C., Shu, S.B., Arentze, T.A., Baker, T., Diehl, K., Donkers, B., Fast, N.J., Häubl, G., Johnson, H., Karmarkar, U.R., et al. (2020). Consumer decisions with artificially intelligent voice assistants. Mark Lett. *31*, 335, https://doi.org/10.1007/s11002-020-09537-5.

Dietvorst, B.J., Simmons, J., and Massey, C. (2014). Understanding algorithm aversion: forecasters erroneously avoid algorithms after seeing them err. Acad. Manage. Proc. 12227, https://doi.org/10.5465/ambpp.2014.12227abstract.

Dietvorst, B.J., Simmons, J.P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. J. Exp. Psychol. Gen. *144*, 114–126.

Elish, M.C. (2019). Moral crumple zones: cautionary tales in human-robot interaction. Engag. Sci. Technol. Soc. *5*, 40, https://doi.org/10.17351/ests2019.260.

Gao, P., Hensley, R., and Zielke, A. (2014). A roadmap to the future for the auto industry. McKinsey Quarterly Oct, 1–11. www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry.

Gefen, D., Karahanna, E., and Straub, D.W. (2003). Trust and TAM in online shopping: an integrated model. MIS Q. 51, https://doi.org/10.2307/30036519.

Geistfeld, M.A. (2017). A roadmap for autonomous vehicles: state tort liability, automobile insurance, and federal safety regulation. Calif. L. Rev. *105*, 1611.

Gerstenberg, T., Ullman T.D., Nagel J., Kleiman-Weiner M., Lagnado D., Tenenbaum j. Lucky or clever? From expectations to responsibility judgments. (n.d.) https://doi.org/10.31234/osf.io/jc72g.

Gerstenberg, T., and Lagnado, D.A. (2010). Spreading the blame: the allocation of responsibility amongst multiple agents. Cognition *115*, 166–171.

Goodall, N.J. (2016). Can you program ethics into a self-driving car? IEEE Spectr. *53*, 28–58, https://doi.org/10.1109/mspec.2016.7473149.

Hidalgo, C.A., Orghiain, D., Canals, J.A., De Almeida, F., and Martín, N. (2021). How Humans Judge Machines (MIT Press).

Lagnado, D.A., Gerstenberg, T., and Zultan, R. (2013). Causal responsibility and counterfactuals. Cogn. Sci. *37*, 1036–1073.

Lambert, F. (2018). Tragic fatal crash in Tesla Model S goes national because - tesla - Electrek. https://electrek.co/2018/05/09/tesla-model-s-fatal-crash-fire-national/.

Lee, J.D., and Moray, N. (1994). 'Trust, self-confidence, and operators' adaptation to automation'. Int. J. Hum. Comput. Stud. *40*, 153–184, https://doi.org/10.1006/ijhc.1994.1007.

Lee, J., and Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics *35*, 1243–1270.

Liggett, T.M. (2012). Interacting Particle Systems (Springer Science & Business Media).

Liu, P., Yang, R., and Xu, Z. (2019). How safe is safe enough for self-driving vehicles? Risk Anal. *39*, 315–325.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents (Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15).

Marcus, G., and Davis, E. (2019). Rebooting AI: building artificial intelligence we can trust. Pantheon.

Matthias, A. (2004). The responsibility gap: ascribing responsibility for the actions of learning automata. Ethics Inf. Technol. *6*, 175–183, https://doi.org/10.1007/s10676-004-3422-1.

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

Ray, S., Chatterjee, K., Majumdar, R., and Ganguly, D. (2021). Voting in watts-strogatz small-world network. Adv. Intell. Syst. Comput. 329–342, https://doi.org/10.1007/978-981-15-7834-2_31.

Schlögl, S., Postulka, C., Bernsteiner, R., and Ploder, C. (2019). Artificial intelligence tool penetration in business: adoption, challenges and fears. In Knowledge Management in Organizations (Springer International Publishing), pp. 259–270.

Shariff, A., Bonnefon, J.F., and Rahwan, I. (2017). Psychological roadblocks to the adoption of self-driving vehicles. Nat. Hum. Behav. *1*, 694–696.

Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Morton, D., and Pavone, M. (2014). Toward a Systematic Approach to the Design and Evaluation of Automated Mobility-On-Demand Systems: A Case Study in Singapore (Road Vehicle Automation), pp. 229–245.

Sun, T.Q., and Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: evidence from public healthcare. Government Inf. Q. *36*, 368–383.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of "small-world" networks. Nature, 440–442, https://doi.org/10.1038/30918.

iScience, Volume *24*

Supplemental information

Blaming automated vehicles in difficult situations

Matija Franklin, Edmond Awad, and David Lagnado

Supplemental Data Items

Item Number	F Causality	F Blame
1	.634	.005
2	11.603***	18.881***
3	2.154	.791
4	7.375**	4.335*
5	.034	.017

Table S1. **Results from individual item ANOVAS between human and machine groups in Study 2, related to Figure 3**. Ten separate ANOVAs which explored the differences in mean causality or blame between groups judging machine or human drivers, for the five different items. For all significant scores, machine drivers received significantly more blame judgments or causal attributions than human drivers p< .05 = *, p< .01 = **, p< .001 = ***]. For non-significant scores in the remaining three items, the same direction either held (i.e., machines were attributed more blame or causality) or the differences were negligible.

Transparent Methods

Study 1

Design

The study used a between-subject design with four experimental groups, which differed in terms of the drivers and mistakes people were judging - human drivers making mistakes in simple situations, AI drivers making mistakes in simple situations, human drivers making mistakes in complex situations, AI drivers making mistakes in complex situations. All groups only differed in the traffic accident items it showed to participants, while all groups showed the same instructions and asked the same additional questions about self-driving car knowledge, and sociodemographic questions. Participants were randomly assigned to the four experimental groups. In response to each item, participants made causal attributions, blamed judgments and were asked whether or not they would make the same mistake, and would someone they know have made the same mistake. The summary of the participants' attributions and judgments - 'Total Blame' and 'Total Cause' - were the study's main dependent variables. The sum of people's response to whether they would make the mistake, or know someone who made a mistake, were used as one of the study's predictor variables.

For each item, participants were additionally asked to write why they made the judgment they made, and also what the driver could have done differently. These qualitative questions served as a way of collecting information that could inform or inspire the design of future studies.

Experimental Procedure

Participants were told that they were going to be presented with six different traffic accident scenarios, for which they were going to make judgments of causality and blame, as well as some additional questions. Participants had the right to leave the study at any point, but responding to all items was mandatory. Informed consent was obtained from all participants before the beginning of the study.

Participants were first shown instructions for what they were going to do in the study. They were then split into one of the four groups. They then responded to the six traffic accident scenarios. Finally, participants were asked what they knew about self-driving cars, and answered six sociodemographic questions. This item order was used because the sociodemographic questions could prime participants' judgments and attributions, an effect known as social priming. Study data will be publicly available on GitHub.

Participants

A power analysis was conducted in order to determine the smallest sample size suitable to detect the effects of an ANOVA. The alpha level set to 0.05 and a power set to 0.8. The estimation indicated that the minimum number of participants had to be 180, with a final sample of 198 achieved.

For the initial study, participants had to be from the UK and above the age of 18. The 198 participants (53% females; median age of 27) were recruited via Prolific.

Measures

Traffic Accident Scenarios

The six scenarios were constructed so that mistakes in complex situations were harder to avoid than mistakes in simple situations. The mistakes always resulted in a crash. Across the four experimental groups the traffic accident scenarios were mostly similar, but differed in two crucial ways. First, for AI groups, the driver was referred to as the "self-driving car", while for human groups, the driver was referred to as the "person" or "driver". Second, there was always one sentence per scenario that differed between simple and complex mistake scenarios. See the Supplementary Information for all items.

Judgments and Attributions

To measure the causal attributions and blame judgments, participants were asked "To what extent did the X cause the crash?" and "To what extent is the X to blame for the crash?", respectively, with X referring to either the human driver or self-driving car. These were made on a scale of 0-100. Participants were also asked "Would have you made the same mistake?" and "Would someone you know have made the same mistake?" to which they could respond yes or nor. For all of the described

questions, participants were asked once per accident item, with six items in total per experimental group.

Qualitative
For every traffic accident scenario, participants were asked the two following qualitative questions - “Why did you make this judgment?” and “What could the X have done differently?” with X referring to either the human driver or self-driving car.

Sociodemographic
Participants were asked to provide information about six sociodemographic factors - age, education level, gender, household income, political views and religious views.

Other
Participants were asked - “What do you know about self-driving cars?”.

Experimental Items

*Human driver experimental groups referred to the driver as “person” and machine driver experimental groups referred to the driver as “self-driving car”. Below are the items for the human driver experimental groups.

I Simple Driving Situations

a. Items

1. A person is driving on a highway. There is another car driving in front. The car in front brakes unexpectedly and the person crashes into it.
2. A car driver is about to enter a multi-lane roundabout. The roundabout is empty. The driver enters and crashes into the centre of it.
3. An ambulance has its sirens on and is trying to get through traffic. The traffic is very light, and a driver in traffic only needs to move to the side to make space. The driver tries to do that but hits a signpost and causes a traffic jam. As a result, the ambulance was too late to save the patient.
4. A person is driving down a quiet road, abiding by the speed limit. Twenty meters ahead of him a pedestrian jumps on the road. The person makes a very sudden turn and crashes into a wall on the side of the road.
5. A person is driving on a highway. There is another car driving in front. The car in front brakes unexpectedly and the person crashes into it.
6. A person is driving down a quiet road, abiding by the speed limit. Twenty meters ahead of him, on the side of the road he sees a person who is about to cross the road. The driver keeps driving and crashes into the person on the road.

b. Questions (repeated for all items)

1. To what extent did the driver cause the crash? 0-100 (slider)
2. To what extent is the driver to blame for the crash? 0-100 (slider)
3. Would have you made the same mistake? i. Yes ii. No
4. Would someone you know have made the same mistake? i. Yes ii. No
5. Why did you make this judgment?
6. What could the driver have done differently?

II Complex Driving Situations

a. Items

1. A person is driving on a highway, it’s raining and the visibility is bad. There is another car driving in front. The car in front brakes unexpectedly and the person crashes into it.
2. A car driver is about to enter a multi-lane roundabout. The roundabout is very busy. The driver enters and crashes into the centre of it.
3. An ambulance has its sirens on and is trying to get through traffic. The traffic is very busy, and a driver in traffic needs to do some manoeuvring in order to make space. The driver tries to do that but he hits a signpost and causes a traffic jam. As a result, the ambulance was too late to save the patient.

4. A person is driving down a quiet road, abiding by the speed limit. Two meters ahead of him a pedestrian jumps on the road. The person makes a very sudden turn and crashes into a wall on the side of the road.
5. A person is driving down the road and the car in front suddenly stops. The person crashes into the car.
6. A person is driving down a quiet road, abiding by the speed limit. Twenty meters ahead of him, on the side of the road he sees an ostrich that is about to cross the road. The driver keeps driving and crashes into the ostrich on the road.

b. Questions (repeated for all items)
*Same as for other experimental groups (See I Simple Driving Situations, b. Questions)

III Additional and Demographic Questions

1. What do you know about self-driving cars?
2. What is your year of birth?
3. What is the highest level of school you have completed or the highest degree you have received?
4. What is your sex?
5. Information about income is very important to understand. Would you please give your best guess? Please indicate the answer that includes your entire household income in (previous year) before taxes.
6. Please indicate your political views from extremely progressive (left) to extremely conservative (right). Where would you place yourself on this scale?
7. Please indicate your religious views from extremely non-religious (left) to extremely religious (right). Where would you place yourself on this scale?

Study 2

Design

The study used a between-subject design with six experimental groups, which differed in terms of the drivers and mistakes people were judging - human drivers making mistakes in simple situations, AI drivers making mistakes in simple situations, human drivers making mistakes in complex situations, AI drivers making mistakes in complex situations, human drivers making mistakes in novel situations, AI drivers making mistakes in novel situations. As with Study 1, all groups only differed in the traffic accident items it showed to participants, while all groups showed the same instructions and asked the same additional (described in Experimental Procedure and Materials) and sociodemographic questions. Participants were randomly assigned to the six experimental groups. In response to each item, participants made causal attributions, blamed judgments and were asked to judge to what extent the described driving situation was difficult and novel. The summary of the participants' attributions and judgments - 'Total Blame' and 'Total Cause' - were the study's main dependent variables. Participants summed judgments on item difficulty and novelty were used to validate the items.

Experimental Procedure

Participants were told that they were going to be presented with five different traffic accident scenarios, for which they were going to make judgments of causality and blame, as well as some additional questions. Participants had the right to leave the study at any point, but responding to all items was mandatory. Informed consent was obtained from all participants before the beginning of the study.

Participants were first shown instructions for what they were going to do in the study. They were then split into one of the six groups. They then responded to the five traffic accident scenarios. The order in which the five items were presented was randomised. The order of these questions for each individual item was also randomised. Participants were then asked whether they wanted to buy a car or self-driving car, and whether they trust other drivers or self-driving cars, depending whether they were in a group that judged human or AI drivers, respectively. They were also asked whether or not they knew a mistake was about to happen in the previous items that described traffic accident scenarios. Finally, to validate one of the experimental items, the participants were asked whether they (or anyone they know) have ever encountered a deer or an ostrich on the road, depending on the experimental group they were in (people judging novel mistakes were asked about the ostrich, while others were asked about the deer).

Participants were then asked some additional questions. First, they were asked to write why they made the previous choices while evaluating the traffic accident items. They were then asked whether they have a driving licence. If they replied yes, they were then asked whether or not they own a car and how many times do they drive per week, on average. Finally, they were asked the same six socio-demographic questions from Study 1. As with Study 1, the item order was used because the sociodemographic questions could prime participants' judgments and attributions. Study data will be publicly available on GitHub.

Participants

A power analysis was conducted in order to determine the smallest sample size suitable to detect the effects of an ANOVA. The alpha level set to 0.05 and a power set to 0.8. The estimation indicated that the minimum number of participants had to be 216, with a final sample of 317 achieved.

For the initial study, participants had to be from the UK and above the age of 18. The 317 participants (36% females; median age of 33) were recruited via MTurk.

Measures

Traffic Accident Scenarios

The five scenarios were constructed so that mistakes in novel and complex situations were harder to avoid than mistakes in simple situations. Two mistakes resulted in a crash, another two in the passengers being late and one resulted in a car crash. Across the six experimental groups the traffic accident scenarios were mostly similar, but differed in two crucial ways. First, for AI groups, the driver was referred to as the "self-driving car", while for human groups, the driver was referred to as the "person" or "driver". Second, there was always one or two sentences per scenario that differed between simple, complex and novel mistake scenarios. See the Supplementary Information for all items.

Judgments and Attributions

To measure the causal attributions and blame judgments, participants were asked "To what extent did the X cause the Y?" and "To what extent is the X to blame for the Y?", respectively, with X referring to either the human driver or self-driving car, and Y referring to the outcome of the mistake. These were made on a scale of 0-100. Participants were also asked "To what extent is the described driving situation novel?" and "To what extent is the described driving situation difficult?" to which they could respond on a 0-100 scale. For all of the described questions, participants were asked once per accident item, with five items in total per experimental group.

Qualitative

After responding to the traffic accident scenarios, participants were asked "Please describe how you made your choices in this HIT".

Sociodemographic

As with Study 1, participants were asked to provide information about six sociodemographic factors - age, education level, gender, household income, political views and religious views.

Other

Participants in human driver experimental groups were asked to reply to "I want to buy a car" and "I trust other drivers" on a 1-7 scale (Strongly disagree - Strongly agree). Participants in AI driver experimental groups were asked to reply to "I want to buy a self-driving car one day" and "I trust self-driving cars" on a 1-7 scale (Strongly disagree - Strongly agree).

Participants in all groups were asked to reply to "In most of the previous stories, as I was reading the story, I knew that a mistake was about to happen" on a 1-7 scale (Strongly disagree - Strongly agree).

Participants in simple and complex mistake experimental groups were asked "Have you, or anyone you know, ever encountered a deer on the road?" to which they could reply "Yes" or "No". Participants in the novel mistake experimental groups were asked "Have you, or anyone you know, ever encountered an ostrich on the road?" to which they could reply "Yes" or "No".

Finally, Participants were asked "Do you have a driving licence?" to which they could reply "Yes" or "No". If they replied with "Yes", they were additionally asked "Do you own a car" to which they could reply "Yes" or "No", as well as "On average, how many times a week do you drive?" to which they needed to respond with a number.

Experimental Items

I Simple Driving Situations

a. Items

1. A person is driving down a road. There is a STOP sign that is clearly visible. The driver doesn't stop and crashes into another car that had priority at that crossing.
2. A person is driving down a quiet road, abiding by the speed limit. Twenty meters ahead of him/her, on the side of the road, he/she sees a deer that is about to cross the road. The driver turns suddenly and crashes into a lamppost.
3. A person driving a car is stopped at a red traffic light. The traffic light turns green. The driver does not move and he/she causes a traffic jam.
4. A person is driving down a two way road. There is a passenger in the back seat. The driver approaches another car that is broken down. There is room for the driver to overtake the broken down car and it is clearly legal to overtake on that part of the road. The driver stops driving and shuts down at a safe distance from the broken down car. The passenger in the car is late for a meeting.
5. A person is driving towards the airport. There is a passenger in the back seat. There is little traffic along the route. This makes it easy for the driver to change to a faster route. However, the driver does not change routes and as a result the passenger is late for his/her flight.

b. Questions (repeated for all items)

*The questions below refer to the questions participants got in response to items that described a crash. The text would change accordingly if the item was referring to a traffic jam or a passenger being late

1. To what extent did the driver cause the crash?
2. To what extent is the driver to blame for the crash?
3. To what extent is the described driving situation novel?
4. To what extent is the described driving situation difficult?

c. Questions (standalone and asked after the five items)

1. I want to buy a car. 1-7 (Strongly Disagree-Strongly Agree)
2. I trust other drivers. 1-7 (Strongly Disagree-Strongly Agree)
3. In most of the previous stories, as I was reading the story, I knew that a mistake was about to happen. 1-7 (Strongly Disagree-Strongly Agree)
4. Have you, or anyone you know, ever encountered a deer on the road? i. Yes ii. No

II Novel Driving Situations

a. Items

1. A person is driving down a road. There is a STOP sign but the S has been scratched off and now the sign reads TOP. The driver doesn't stop and crashes into another car that had priority at that crossing.
2. A person is driving down a quiet road, abiding by the speed limit. Twenty meters ahead of him/her, on the side of the road, he/she sees an ostrich that is about to cross the road. The driver turns suddenly and crashes into a lamppost.
3. A person driving a car is stopped at a red traffic light. The traffic light turns green, but the green glass is broken so the light appears white. The driver does not move and he/she causes a traffic jam.
4. A person is driving down a two way road. There is a passenger in the back seat. The driver approaches another car that is broken down. There is room for the driver to overtake the broken down car but the double white line indicates that it is illegal to overtake on that part of

the road. The driver stops driving and shuts down at a safe distance from the broken down car. The passenger is late for a meeting.

5. A person is driving towards the airport. There is a passenger in the back seat. The radio announces that the concert in a nearby stadium has finished earlier than planned and there will be a lot of traffic along the route. This will soon make it tricky for the driver to change to a faster route. However, the driver does not change routes and as a result the passenger is late for his/her flight.

b. Questions (repeated for all items)
*Same as for other experimental groups (See I Simple Driving Situations, b. Questions)

c. Questions (standalone and asked after the five items)
*Same as for other experimental groups (See I Simple Driving Situations, c. Questions) with one exception - question 4:

1. Have you, or anyone you know, ever encountered a ostrich on the road? i. Yes ii. No

III Complex Driving Situations
a. Items

1. A person is driving down a road. There is a STOP sign but it is foggy and the visibility is very bad. The driver doesn't stop and crashes into another car that had priority at that crossing.
2. A person is driving down a quiet road, abiding by the speed limit. Two meters ahead of him/her, on the side of the road, he/she sees a deer that is about to cross the road. The driver turns suddenly and crashes into a lamppost.
3. A person driving a car is stopped at a red traffic light. The traffic light turns green, but only for one second due to a fault with the traffic light. The driver does not move and he/she causes a traffic jam.
4. A person is driving down a two way road. There is a passenger in the back seat. The driver approaches another car that is broken down. There is little room for the driver to overtake the broken down car with incoming traffic but it is clearly legal to overtake on that part of the road. The driver stops driving and shuts down at a safe distance from the broken down car. The passenger is late for a meeting.
5. A person is driving towards the airport. There is a passenger in the back seat. It is rush hour and there is a lot of traffic along the route. This makes it tricky for the driver to change to a faster route. However, the driver does not change routes and as a result the passenger is late for his/her flight.

b. Questions (repeated for all items)
*Same as for other experimental groups (See I Simple Driving Situations, b. Questions)

c. Questions (standalone and asked after the five items)
*Same as for other experimental groups (See I Simple Driving Situations, c. Questions)

IV Additional and Demographic Questions

1. Please describe how you made your choices in this HIT.
2. Do you have a driving licence? i. Yes ii. No [If "Yes" is Selected - 1. Do you own a Car? i. Yes ii. No; On average, how many times a week do you drive?]
3. What is your year of birth?
4. What is the highest level of school you have completed or the highest degree you have received?
5. What is your sex?
6. Information about income is very important to understand. Would you please give your best guess? Please indicate the answer that includes your entire household income in (previous year) before taxes.
7. Please indicate your political views from extremely progressive (left) to extremely conservative (right). Where would you place yourself on this scale?
8. Please indicate your religious views from extremely non-religious (left) to extremely religious (right). Where would you place yourself on this scale?