







# The Blame Game: Double Standards Apply to Autonomous Vehicle Accidents

Qiyuan Zhang<sup>(✉)</sup> , Christopher D. Wallbridge , Dylan M. Jones ,  
and Phil Morgan 

Cardiff University, Cardiff CF10 3AT, UK  
zhangq47@cardiff.ac.uk

**Abstract.** Who is to blame when autonomous vehicles are involved in accidents? We report findings from an online study in which the attribution of blame and trust were measured from 206 participants who studied 18 hypothetical vignettes portraying traffic incidents under different driving environments. The focal vehicle involved in the incident was either controlled by a human driver or autonomous system. The accident severity also varied from near miss, minor accident to major accident. Participants applied double standards when assigning blame to humans and autonomous systems: an autonomous system was usually blamed more than a human driver for executing the same actions under the same circumstances with the same consequences. These findings not only have important implications to AI-related legislation, but also highlight the necessity to promote the design of robots and other automation systems which can help calibrate public perceptions and expectations of their characteristics and capabilities.

**Keywords:** Trust · Blame · Autonomous vehicles · Liability · Automobile accident · Human-robot interaction

## 1 Introduction

Improved road safety and alleviated traffic congestion, among others, are likely to flow from the adoption of autonomous vehicle (AV) technology but many challenges need to be overcome before they are widely accepted. In addition to the reliability issues associated with the technology itself, ethical and legal implications need to be considered when we determine the way in which we should anticipate, stipulate and appraise the decisions and behaviors of an AV. One of the most difficult practical challenges facing legislators and policy makers is how responsibilities/liabilities should be distributed among different parties following accidents in which an AV is involved [1, 3, 5–7].

The legal infrastructure in most countries assumes that the human driver possesses full control of a non-autonomous vehicle and therefore full responsibility for the safety of all its passengers and other road users. This legal framework is less appropriate for semi- or fully autonomous vehicles where the driver/user of the car relinquishes partial or complete control of the vehicle to an automated system [7]. This transition of control means a shift of responsibility from the driver of the vehicle to a set of entities spanning

the car manufacturer, software programmer and government. But it is not always easy to pinpoint the location where one party's responsibility should end and another starts when the duty of driving is shared between the vehicle and a person [2, 6]. Even with fully autonomous vehicles where the role of the user is reduced to that of a passenger several legal issues still remain. For example, how should the Artificial Intelligence (AI) which operates the vehicle be treated as a legal entity? Should it be treated as a legal person or should we divide responsibility among its progenitors (e.g., manufacturers, software developers, etc.)? The answers to these questions will have profound impact in shaping the automobile industry as well as society.

Despite a recent surge in human factors literature on autonomous driving, empirical research is relatively sparse on the topic of responsibility and blame attribution in relation to AVs, especially at high levels of automation (for limited example, see [2, 8]). The aim of the current study is to inform debate by studying observers' intuitions and attitudes about liability for automobile accidents involving fully autonomous vehicles.

## 1.1 Hypotheses

We proposed that people apply double standards to autonomous and human- driven vehicles when making judgements of blame and trust and that these judgements are also a function of outcome severity. Our key hypotheses were:

- H1** Fully autonomous cars will be blamed more than manually driven cars for the same action in a traffic incident.
- H2** Human drivers will be more trusted than autonomous systems after being involved in a traffic incident.
- H3** Blame on AVs will increase as the accident severity increases.
- H4** Trust in AVs will diminish as the accident severity increases.

## 2 Methodology

206 participants were recruited and paid through *Prolific Academic*<sup>1</sup>. Pre-screening criteria dictated that all were believed to be UK residents, over 18 years old, with normal or corrected-to-normal vision.

The study adopted a mixed 2 (Operator) X 3 (Outcome Severity) X 6 (Scenario) design. Each participant was presented with six scenarios in a random order. Outcome Severity was manipulated as a repeated-measure variable at three levels: Near miss (an accident is narrowly avoided), Minor Accident (property damage but no personal injury) and Major Accident (property damage and personal injury). Participants experienced all three outcomes for each scenario in a counterbalanced order. The factor Operator was between-participant: Half the participants were told that the target vehicles in the scenarios were controlled by fully autonomous systems and the other half that they had human drivers.

The key dependent variable of interest was attribution of blame: The extent to which the driver/controller of a vehicle should be blamed for the outcome. The level of trust

<sup>1</sup> <https://www.prolific.co/>.

was measured via a judgment about whether the driver/controller of the vehicle could be relied on to safely operate a vehicle in the future.

There were six scenarios: (i) a child running out from between parked cars; (ii) a pedestrian crossing the road; (iii) passengers crossing in front of a stopped bus; (iv) a deer jumping onto the road; (v) a second car pulling out of an intersection; and (vi) a tree falling on the road. Each scenario comprised a sequence in which Part A described the emergency situation and Part B described the actions of Vehicle X as well as the incident's outcome.

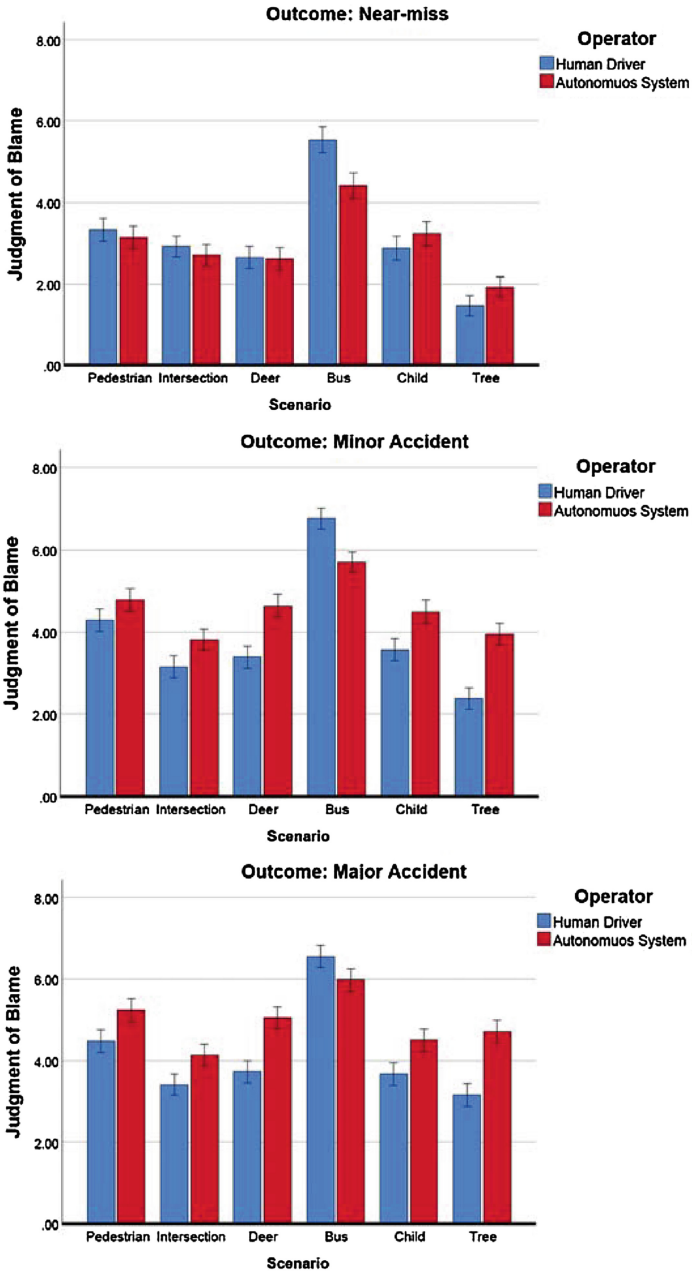
Each part consisted of a textual description and a pictorial illustration. For each Part A there could be one of three possible outcomes in Part B. For example, for the scenario 'Child' (i, above), Part A read 'You are a passenger riding in Vehicle X, which is driving slowly down a street with parked cars on either side. A child runs out from between two cars.' Part B with Near miss outcome read: 'Vehicle X swerves to the left to avoid hitting the child. It narrowly misses the child. No collision occurs.' Part B with Minor Accident outcome read: 'Vehicle X swerves to the left. It misses the child but crashes into one of the parked cars in the process. No personal injury is caused to anyone.' Part B with Major Accident outcome is similar to Part B with Minor Accident but the last sentence was replaced by 'You suffer minor injuries.' Questions following every Part B measured post-incident blame and trust on an 11-point scale.

## 3 Results

### 3.1 Post-incident Blame

Participants' ratings of the post-incident blame were cast into a 2 (Operator) X 3 (Outcome Severity) X 6 (Scenario) mixed ANOVA, which revealed a significant main effect of Scenario ( $F(5, 1020) = 63.781, p < .001, \eta^2 = .238$ ) and Outcome Severity ( $F(2, 408) = 133.800, p < .001, \eta^2 = .396$ ). The level of blame attribution varied from scenario to scenario: The more severe the outcome, the greater the blame. There was also a marginally significant main effect of Operator ( $F(1, 204) = 3.204, p = .075, \eta^2 = .015$ ). In most scenarios there was a tendency to blame an autonomous system more than a human driver. Importantly, this effect was moderated by Outcome Severity, which was evidenced by a significant two-way interaction between Outcome Severity and Operator ( $F(2, 408) = 12.724, p < .001, \eta^2 = .059$ ) - An autonomous system received more blame than a human driver only when the outcome of the incident was consequential (i.e., minor or major accidents, not near miss).

The effect of Operator was not consistent across scenarios, confirmed by a significant two-way interaction between Scenario and Operator ( $F(5, 1020) = 8.278, p < .001, \eta^2 = .039$ ). In five out of six scenarios, participants assigned significantly greater blame to an autonomous system than a human driver when the outcome was consequential. But this pattern was reversed in the Bus Scenario: The human driver was blamed more than the autonomous system, which indicates that the direction of discrimination against autonomous vehicles and human drivers is to some degree context dependent (Fig. 1).



**Fig. 1.** Mean ratings of post-incident blame across all scenarios and outcomes (Error bars =  $\pm 1$  SE)

### 3.2 Post-incident Trust

Ratings of post-incident trust in the driver/operator of the target vehicle (Vehicle X) display a mirror image of the ratings of blame. A 2 (Operator) X 3 (Outcome Severity) X 6 (Scenario) mixed ANOVA revealed a significant main effect of Scenario ( $F(5, 1020) = 52.129, p < .001, \eta^2 = .204$ ) and Outcome Severity ( $F(2, 408) = 187.428, p < .001, \eta^2 = .479$ ). Like blame, ratings of post-incident trust varied from scenario to scenario and diminished in magnitude as the severity of the outcome increased. The main effect of Operator was also significant ( $F(1, 204) = 33.990, p < .001, \eta^2 = .143$ ). Participants trusted the driver of the vehicle less if it was an autonomous system than if it was a human. But again, the magnitude of this effect was found to be dependent on both scenario ( $F(5, 1020) = 6.120, p < .001, \eta^2 = .029$ ) and outcome ( $F(2, 408) = 6.501, p = .002, \eta^2 = .031$ ). Like blame, the effect of Operator on trust was more pronounced after consequential outcomes than after near misses. Secondly, the magnitude of this effect was found to be smaller in the Bus scenario, but not reversed like the ratings of blame. Together these results suggest that blame and trust are closely associated constructs and post-incident trust is at least partly informed by blame.

## 4 Discussion

The study was successful in showing systematic effects of the type of vehicle and outcome severity on blame attribution and trust after an accident. Consistent with previous research (e.g., [9]), the level of blame attributed to the operator was positively related to accident severity, with near misses eliciting the lowest level of blame while major accidents received the highest level. These results are perhaps not surprising due to the fact that negative emotions have been found to be a major contributing factor in the attribution of blame (see [4] for a review) and severe outcomes, especially those involving personal injuries, are more readily to provoke emotional reactions than less severe outcomes.

The distinctive feature of our results is the evidence of double standards for autonomous systems and human drivers when ascribing blame. This discrimination was the most evident when the traffic incident was consequential (that is, not a near miss). Moreover, the direction of this discrimination was not universal: In five out of six scenarios the autonomous systems received more blame than human drivers, but this pattern was reversed in the bus scenario, where the human driver was blamed to a greater extent than the autonomous system. This is in sharp contrast with that of some previous studies (e.g., [2, 8]) showing that humans are judged more harshly than autonomous vehicles. We would argue that the current study has a wider set of scenarios and has better control of variables than other studies, which leads us to suggest that our findings are more accurate and comprehensive. Certainly, this aspect of our findings will need to be confirmed and extended.

But why the anomalous finding in the Bus scenario? We suggest that the normative expectation is that autonomous systems are expected to outperform humans in reacting to an emergency whereas humans perform better in anticipating dangers based on event cues and taking proactive measures. In most of our scenarios, the emergency occurs very suddenly and unpredictably. For example, they usually feature a pedestrian, a vehicle, an

animal or an object suddenly jumping/falling in front of the vehicle with the driver/system needing to respond quickly to avoid a crash. In these settings, computers might be expected to outperform their human counterparts, due to the perception that they have more advanced sensory systems and faster computing/processing speed. This heightened expectation for performance in reacting to danger, justified or not, might have resulted in AVs getting more blame when a crash does happen. In comparison, in our outlier, the bus scenario, there are event cues (e.g., a stationary bus at a bus stop unloading passengers) to help the human driver predict what is about to occur (e.g., the likelihood that some disembarking passengers will want to cross the road). Hence providing an opportunity for the driver of Vehicle X to be proactive and carry out preventative measures (e.g., slowing the car down). This is despite the fact that the visibility of the pedestrians within our scenario is obscured by the bus. The typical observer might not expect a machine to possess the same capability as the human for drawing this type of causal connection (e.g., predicting what some of the passengers might do without being able to 'see' through the bus). This speculation is supported by the fact that the blame levels in this scenario were generally higher than other scenarios, indicating that the events were more foreseeable and hence more preventable than in other scenarios and yet more foreseeable for a human driver. This proposition that the nature of the discrimination between humans and machines is moderated by the perceived foreseeability of the emergency situation needs to be formally tested by future research.

Our study also revealed that post-incident blame on a particular autonomous system can inform trust in the same autonomous system. This is supported by the fact that trust ratings displayed a reciprocal pattern to that of the ratings of blame across all conditions. This suggests that not only do people apply double standards to humans and autonomous systems in retrospective judgment of blame, this discrimination can also be carried over to affect their future decisions regarding the adoption of different modes of transport, which highlights the necessity to promote design principles that facilitate the calibration of the public's expectations with regard to the operating capabilities of AVs.

## 5 Conclusion

In contrast with the big strides being made in the development of AI-related technologies (e.g., deep learning, quantum computing, etc.), scant attention has been given to the human consequences of these developments. Distrust and fear of AI, combined with the uncertainties associated with regulations, laws and ethics when interacting with such technologies, have become major hindrances for AI adoption, especially in safety-critical domains such as transport, medicine, and security. Before societies can fully embrace AI, an important theoretical and practical question needs to be answered: Who is to blame when things go wrong?

AI-enabled autonomous systems such as self-driving cars already possess the ability to make decisions and carry them out independently without human supervision or approval. The pace of improvement in this technology is rapid. Yet answers to the question of who should be held accountable for their actions has hardly received any systematic attention.

Our study has taken a step towards developing a robust experimental paradigm that can be used to explore a wide range of phenomena. For example, future research

could use a variety of scenarios and applications with different experimental stimuli (e.g., simulations of higher fidelity) and explore such factors as the moderating effect of standpoints (e.g., victim of a car crash versus bystanders). Only then we can begin to synthesize conceptual models in order to better predict public perceptions of autonomous vehicles and embody them in the design of vehicles and the operational framework of future transport systems.

**Acknowledgments.** The research was funded through the ESRC Project ES/T007079/1 Rule of Law in the Age of AI: Principles of Distributive Liability for Multi-Agent Societies and is part of a larger project on the same topic supported by ESRC (ES/T007079/1) and JST with collaborators at the Universities of Kyoto, Osaka and Doshisha. We would like to thank our collaborators Prof. Tatsu Inatani and Prof. Minoru Asada for their contribution to this work.

## References

1. Anderson, J.M., Kalra, N., Wachs, M.: Liability and regulation of autonomous vehicle technologies. RAND Corporation, Berkeley, CA (2009). <http://www.rand.org/pubs/externalpublications/EP20090427.html>
2. Awad, E., et al.: Drivers are blamed more than their automated cars when both make mistakes. *Nat. Hum. Behav.* **4**(2), 134–143 (2020)
3. Bellet, T., et al.: From semi to fully autonomous vehicles: new emerging risks and ethico-legal challenges for human-machine interactions. *Transp. Res. Part F: Traffic Psychol. Behav.* **63**, 153–164 (2019)
4. Feigenson, N., Park, J.: Emotions and attributions of legal responsibility and blame: a research review. *Law Hum. Behav.* **30**(2), 143 (2006)
5. Gurney, J.K.: Sue my car not me: products liability and accidents involving autonomous vehicles. *U. Ill. JL Tech. Pol'y*, p. 247 (2013)
6. Hancock, P.: Some pitfalls in the promises of automated and autonomous vehicles. *Ergonomics* **62**(4), 479–495 (2019)
7. Ilková, V., Ilka, A.: Legal aspects of autonomous vehicles—an overview. In: 2017 21st International Conference on Process Control (PC), pp. 428–433. IEEE (2017)
8. Li, J., Zhao, X., Cho, M.J., Ju, W., Malle, B.F.: From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report. SAE Technical Paper (2016)
9. Pöllänen, E., Read, G.J., Lane, B.R., Thompson, J., Salmon, P.M.: Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system. *Ergonomics* **63**(5), 525–537 (2020)