Transportation Research Part A 179 (2024) 103887

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents☆

Qiyuan Zhang[a,c,d,e], Christopher D. Wallbridge[b,c], Dylan M. Jones[a,c,d,1], Phillip L. Morgan[a,c,d,e,f,*]

[a] *School of Psychology, Cardiff University, 70 Park Place, Cardiff CF10 3AT, UK*
[b] *School of Computer Science & Informatics, Abacws, Senghennydd Road, Cardiff CF24 4AG, UK*
[c] *Cardiff University Centre for AI, Robotics and Human-Machine Systems (IROHMS), United Kingdom*
[d] *Human Factors Excellence (HuFEx) Research Group, United Kingdom*
[e] *Cardiff University Digital Transformation Innovation Institute (DTII), United Kingdom*
[f] *Luleå University of Technology - Psychology, Division of Health, Medicine & Rehabilitation, Sweden, Regnbagsallen 5, 977 54 Lulea, Sweden*

ARTICLE INFO

Keywords:
Autonomous driving
Blame attribution
Liability
Trust in automation
Human-robot interaction
Artificial intelligence

ABSTRACT

Road accidents involving autonomous vehicles (AVs) will not only introduce legal challenges over liability distribution but also generally diminish the public trust that may make itself manifested in slowing the initial adoption of the technology and call into question the continued adoption of the technology. Understanding the public's reactions to such incidents, especially the way they differentiate from conventional vehicles, is vital for future policy-making and legislation, which will in turn shape the landscape of the autonomous vehicle industry. In this paper, intuitive judgments of blame and trust were investigated in simulated scenarios of road-traffic accidents involving either autonomous vehicles or human-driven vehicles. In an initial study, five of six scenarios showed more blame and less trust attributed to autonomous vehicles, despite the scenarios being identical in antecedents and consequences to those with a human driver. In one scenario this asymmetry was sharply reversed; an anomaly shown in a follow-up experiment to be dependent on the extent to which the incident was more likely to be foreseeable by the human driver. More generally these studies show—rather than being the result of a universal higher performance standard against autonomous vehicles—that blame and trust are shaped by stereotypical conceptions of the capabilities of machines versus humans applied in a context-specific way, which may or may not align with objectively derived state of affairs. These findings point to the necessity of regularly calibrating the public's knowledge and expectation of autonomous vehicles through educational campaigns and legislative measures mandating user training and timely disclosure from car manufacturers/developers regarding their product capabilities.

☆ This work was funded by the ESRC-JST project: Rule of Law in the Age of AI: Principles of Distributive Liability for Multi-Agent Societies (Morgan, UK Principal Investigator).

* Corresponding author at: School of Psychology, Cardiff University, UK.
E-mail address: morganphil@cardiff.ac.uk (P.L. Morgan).

[1] Deceased.

https://doi.org/10.1016/j.tra.2023.103887
Received 23 April 2022; Received in revised form 2 November 2023; Accepted 6 November 2023
Available online 22 November 2023

1. Introduction

Even when all road vehicles are autonomous, there will still be accidents, some of which will result in considerable harm. Judgments of blame and liability will be shaped both by objective evidence and beliefs about the role played by the autonomous vehicle (AV) in determining the outcome, which may in turn be influenced by expectations of the operating capabilities of the autonomous agency —both specific and general—and how they contrast with those of a human operator (Furlough et al., 2019; T. Kim & Hinds, 2006; J. Lee et al., 2021; J. D. Lee & See, 2004; Madhavan et al., 2006; Muir, 1994). Formal legal ascription of responsibility, mostly aided by expert opinions — e.g., accident investigators, legal professionals, among others, will have huge impacts on stakeholders (e.g., manufacturers, insurers, consumers, etc.) and perhaps even the longevity of the technology (Bellet et al., 2019). Informal human judgments of blame, on the other hand, such as those formed through first-hand experience or media reports, will be driven more by intuition and emotions (Malle et al., 2012). Nonetheless, non-expert opinions may have the power to feed into legal proceedings, e.g., through jury verdicts in courtrooms. They will also influence public trust (de Visser et al., 2018; P. H. Kim et al., 2009) and in turn potentially determine the likelihood of autonomous vehicle adoption or continued use (Adnan et al., 2018; Choi & Ji, 2015; Xu et al., 2018). Understanding the characteristics of these judgments is therefore important to legislators and policy makers in ensuring the viability of autonomous vehicle technologies, as well as to citizens and the public who should form the majority of adopters.

Yet, while much research has been devoted to the factors of trustworthiness and acceptance of autonomous vehicles in everyday ordinary usage (e.g., Abe et al., 2015; Gkartzonikas & Gkritza, 2019; Gold et al., 2015; Hartwich et al., 2018; Nordhoff et al., 2019; Schaefer & Straub, 2016; Waytz et al., 2014), there is a dearth of research on how observers attribute blame in autonomous vehicle accidents along with its consequences for trust in autonomous systems more generally. Using textual and graphic vignettes describing traffic incidents and their antecedents, the work reported here examines how blame and trust are shaped by the knowledge of whether the vehicle is driven by a human driver (HD) or by an autonomous system (AS).

Legally, in many jurisdictions around the world, a human driver is expected to take full responsibility for a conventional vehicle while driving. For example, the Vienna Convention on Road Traffic (1968) stipulated that (human) drivers should retain control of the vehicle at all times and in most European states, (human) drivers of cars with assistive driving technologies are still required to monitor the system and can be charged with negligence in the event of an accident while the assistive system is in control. But this rule becomes progressively less plausible as vehicles become increasingly automated (Anderson et al., 2014; Gurney, 2013; Ilková & Ilka, 2017; Księżak & Wojtczak, 2022; Paret et al., 2022; Pattinson et al., 2020). For instance, in soon-to-be-realised Level-3 automated vehicles, human drivers will no longer be required to actively engage in the driving task, assuming instead the role of a failsafe mechanism, intervening only at the behest of the autonomous vehicle's computer (SAE International, 2018). However, in the case of a transition of control, it becomes problematic to pin-point the moment at which humans can reasonably be assumed to have full control of the vehicle and hence become legally responsible for the behaviour of the autonomous vehicle (see Elish (2019) for a discussion on "moral crumple zones"). Nor are fully-autonomous vehicles without issues of responsibility. For instance, what level of competence should be expected of an autonomous system? How can responsibility be divided among the user, the manufacturer and the programmer who writes the AI algorithm? Clearly, obstacles involving responsibility and liability lie in the path of autonomous vehicle adoption (Bellet et al., 2019; Hancock, 2019).

These difficulties are further exacerbated by the lack of explainability of the decision process of the autonomous system. While a human driver's state of mind and physical capacity can be inferred from physical evidence and testimony, the autonomous system's causal role in an accident may be particularly inscrutable, especially if its computing system is based on 'deep learning' (Doshi-Velez & Kim, 2017; Gilpin et al., 2019; Ly & Akhloufi, 2021; Zablocki et al., 2021). At the same time, the causal chain of events as represented by the autonomous system's actions may be so complex as to make a mechanistic attribution of blame infeasible, even for human experts. However, at least in the foreseeable future, due to the potential personal and social impacts of the tort cases relating to these accidents, the legal judgments will still be beyond the purview of artificial intelligence and primarily lie with humans (e.g., the jury, the judge, etc.). It follows that it is important to understand the human cognitive processes in formulating blame and understanding how this shapes trust in autonomous systems.

Blame is a moral judgment directed at an agent whose behaviours are deemed to have deviated from a norm (Scanlon, 2008). It is not only based on an understanding of the agent's contribution to a negative outcome but also on their reasoning, intentionality, obligation and capacity to prevent the outcome (Malle et al., 2012). The question as to whether inanimate objects like tools and machines are devoid of their own morality or intentions is still a topic of philosophical debate (Hornborg, 2021; Latour, 1993, 1996). It is argued here that a piece of technology could be at least infused with the morality and intentions of their makers/designers. Research in the area of human–machine interaction suggests that people treat computers as though they were social actors (Nass et al., 1994) and humans do attribute blame to autonomous systems (e.g., robots) in collaborative tasks (Furlough et al., 2019) which is a crucial factor of trust restoration after system failures (de Visser et al., 2018).

Trust is another psychological construct that is essential for human acceptance, adoption and continued use of automation systems (J. D. Lee & See, 2004). The role of trust has been examined extensively (e.g. within social psychology and other disciplines) in the context of inter-personal relationships. It has been considered by many theorists of virtue ethics to be a virtue to ensure the flourish of human society which depends vitally on inter-personal collaboration (D'Olimpio, 2018; MacIntyre, 2013; Shionoya, 2001). Evidence has shown that trust is a critical factor in organising economic activity such as commodity production (Nyhan, 2000) and market transactions (R. Morgan & Hunt, 1994). Despite its popularity in academic research, there is no consensus on the definition of trust due to its complex and multi-faceted nature. One of the most cited definitions was given by Mayer et al (1995, p. 712): "willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party". This definition was not developed for human interactions with automation and technologies but it denotes two important components of trust: the relinquish of control and the willingness to assume risk. This stays true when trust is discussed in a the context of human-automation interactions, where trust can be seen as the human operator's intention to concede some or all of the control to an automated system while engaging in little or no supervision. By doing so, the human operator assumes the risk that the automation will not perform to the expected standard. It has been argued that the objective of an automation system design is to facilitate the *calibration* of the operator trust – that is, to achieve a trust level that is proportionate to the capability of the automation system since under- or over-trust could lead to disuse, misuse and abuse of automation (J. D. Lee & See, 2004; Parasuraman & Riley, 1997; Sheridan, 2019).

Until now, human-centred research on autonomous vehicles has focused on ways in which systems can be designed, including ways they can be configured to promote their acceptance and trustworthiness (e.g., Adnan et al., 2018; Choi & Ji, 2015; Forster et al., 2017; Gkartzonikas & Gkritza, 2019; P. L. Morgan et al., 2019; Nastjuk et al., 2020; Panagiotopoulos & Dimitrakopoulos, 2018; Xu et al., 2018) or the safety in handover between a human driver and autonomous system (e.g., Eriksson & Stanton, 2017; Lorenz et al., 2014; Merat et al., 2014; Merat & Jamson, 2009; P. L. Morgan et al., 2018). While acknowledging its significant contribution to the development of autonomous driving technology—to its accessibility, usability and safety—this research was not intended to provide an understanding how blame is distributed in the event of accidents and how it may affect trust. Lee et al (2021) examined the effect of automation failure on trust in automated vehicles but their investigation was restricted to scenarios where a malfunction or an overload of system capacity has resulted in a need for human take-over, not an accident. In the few other cases where accidents have been studied, the emphasis has been on decision making about accident antecedents, with the goal of understanding moral and ethical reasoning about autonomous vehicles, on what observers consider to be the most ethical decision once an accident becomes unavoidable — examples include the 'trolley problem' (Awad et al., 2018; Bonnefon et al., 2016, 2019; Geisslinger et al., 2021; Kallioinen et al., 2019) — rather than on the attribution of blame and its effect on trust after the accident has taken place.

Research on blame attribution in which human drivers share the task of driving with an autonomous system reveals an inconsistent picture. Some studies show that the distribution of blame is proportionate to the perceived control that either party possesses over the outcome (Bennett et al., 2020; McManus & Rutchick, 2019; Pöllänen et al., 2020), human drivers receiving more blame than the autonomous system even though both parties contributed errors to the accident (Awad et al., 2020). Other work showed autonomous systems more to blame and responsible and their actions less acceptable, than that of their human driver counterpart when either the human driver or the autonomous system was depicted to be solely responsible for the crash of a semi-autonomous vehicle (Liu & Du, 2021). However, the discrepancies in blame attribution in this case might be the result of differing descriptions of the cause: For the autonomous system it was described as a 'malfunction', whereas in the human driver condition it was 'inattentiveness'.

Studies of accidents with fully autonomous systems show that people apply inconsistent standards when making a judgment of blame on these systems compared to on human drivers.. A scenario in which both parties are partly at fault—a jaywalking pedestrian is hit by a car—showed that the blame on the driver of the car was higher if it was driven by an AI than by a human (Hong, 2020). The inclination to assign more blame to autonomous systems may reflect the general perception that autonomous systems should be more accomplished in driving than human drivers (Penmetsa et al., 2019; Schoettle & Sivak, 2014). The study of human behaviour more generally has shown that the violation of norms or expectations (e.g., regarding a party's capacity to prevent a negative event) is an important component of counterfactual thinking (e.g., "Someone could/should have done something") as well as blame attribution (Kahneman & Miller, 1986; Malle et al., 2012, 2014; Roese & Olson, 1997; Sanna & Turley, 1996). Moreover, the repair of trust in a party after a breakdown in a human-to-human collaboration is determined by the extent to which the failure is attributable to that party (P. H. Kim et al., 2009). That blame is associated with the violation of prior expectations of performance underpins the hypothesis that has guided the current work.

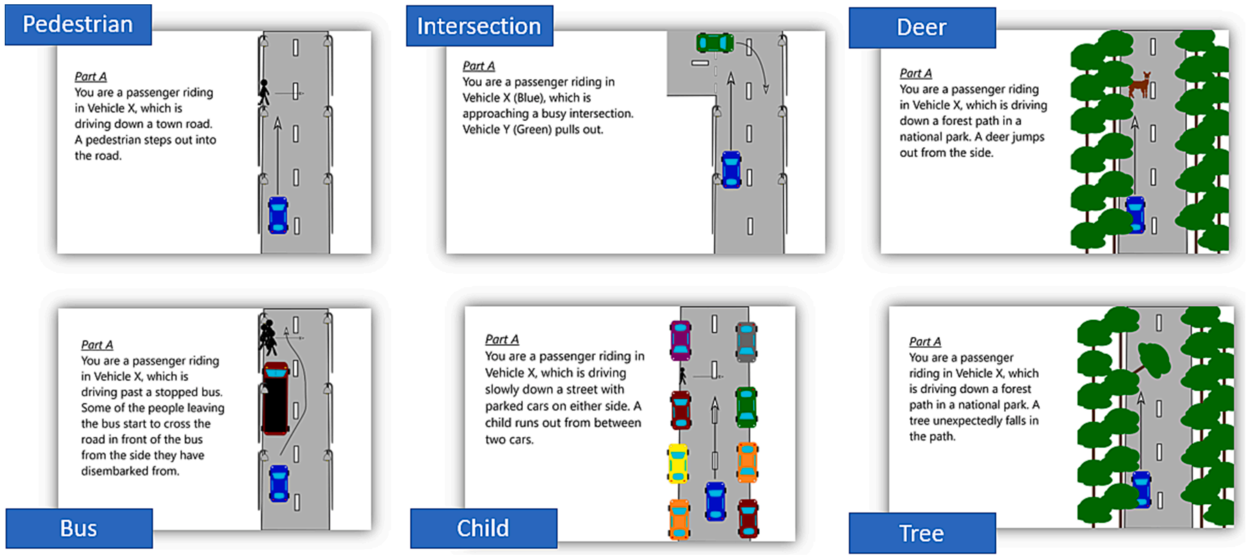
Human expectations possess the property of being complex and contingent. While machines are judged to surpass humans in speed and power, humans are believed to be superior in improvising, inductive thinking and judgments based on prior experience (de Winter & Hancock, 2015; Fitts et al., 1951; Hancock et al., 2019). This suggests that in conditions in which human attributes are still perceived to be advantageous, autonomous systems might be more readily forgiven for failing to avert an accident. Empirical evidence for this is scant, however. One study did not find any evidence for such adaptive behaviour: When the difficulty of the scenario increased, blame for human drivers and autonomous systems decreased equivalently (Franklin et al., 2021). In this connection it is nevertheless important to acknowledge that *difficulty* is a multi-faceted construct that is manifested along many dimensions (familiar *vs.* unfamiliar, likely *vs.* unlikely, low-density traffic *vs.* high-density traffic, low visibility *vs.* high visibility, and so forth) not all of which will act analogously in shaping blame. Arguably therefore, *difficulty* is too gross a variable to capture the highly contingent and intricate mechanisms that shape human expectations.

1.1. The present experiments

We report two experiments demonstrating the adaptive nature of perceived capabilities of autonomous vehicles and their influences on post-accident attitudes, in which volunteer participants, acting alone using a web-based application, made judgments of blame and trust relating to autonomous vehicles or human-driven vehicles featured in narratives in the form of a graphic and verbal vignettes describing road traffic incidents of different levels of severity.

Experiment 1[2] used a set of six incident scenarios (Fig. 1). Inherent in each was some degree of ambiguity about responsibility for

[2] A small sub-set of the results from Experiment 1 were presented at the 12th The International Conference on Applied Human Factors and Ergonomics (2021).



Part A
You are a passenger riding in Vehicle X, which is driving down a town road. A pedestrian steps out into the road.

Part A
You are a passenger riding in Vehicle X (Blue), which is approaching a busy intersection. Vehicle Y (Green) pulls out.

Part A
You are a passenger riding in Vehicle X, which is driving down a forest path in a national park. A deer jumps out from the side.

Part A
You are a passenger riding in Vehicle X, which is driving past a stopped bus. Some of the people leaving the bus start to cross the road in front of the bus from the side they have disembarked from.

Part A
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 A
You are a passenger riding in Vehicle X, which is driving down a forest path in a national park. A tree unexpectedly falls in the path.

Fig. 1. Part A of all six scenarios (Full coloured version available in the digital copy of this paper).

the accident as between the vehicle in question and one or more third parties.

In pilot work, each scenario was iteratively refined to maximise ease and consistency of understanding across participants, consistent with the plan to use un-monitored volunteer participation in the experiment proper. The resulting set of scenarios was relatively broad in coverage of plausible incidents, varying in likelihood, embodying broad equivalence in intelligibility and duration while capturing a degree of variety of agency in third-party involvement, comprising animal, human and institutional agents. However, it is fair to say also the set cannot be said to be exhaustive or even taxonomically coherent. It was designed as a first step to establish consistency of judgement of autonomous vehicles across settings.

By changing details at the narrative, the severity of outcome was manipulated either as a near-miss, minor or major accident. This allowed scrutiny of the role of motivational factors in attribution, particularly the tendency to increase attribution of blame to a party in a negative event as its severity increases, as a way of mentally distancing themselves from the adversity (Fiske & Taylor., 1991; Robbennolt, 2000). The focal vehicles in the scenarios were said to have been driven either by autonomous systems or human drivers, which allowed a direct comparison of the attitudes towards two types of operators on every level of outcome severity.

To anticipate, Experiment 2 examined in detail one anomalous scenario from Experiment 1 (the bus scenario), in which the general trend found in the rest of the set was reversed. We speculated that in this scenario the accident was more readily foreseen and that this capacity to anticipate outcomes was perceived as more of a property of the human than the autonomous vehicle. To put this conjecture to the test, in Experiment 2 we investigated how the strength of causal cues moderated blame and trust using four scenarios adapted from those of Experiment 1.

2. Experiment 1

Experiment 1 was conducted with several working hypotheses. Specifically, we predicted that fully autonomous systems (ASs) would be *blamed* more and *trusted* less than human drivers (HDs) for the same incident outcome and moreover that *blame* on ASs would increase, and trust diminish as outcome severity increased. In addition, we expected that overall, blame and trust on/in the operator of the vehicle involved in a traffic incident would be negatively correlated.

2.1. Method

2.1.1. Participants

206 participants were recruited through the web-based crowd-source *Prolific Academic* to take part in the study and paid £3.75 (GBP) each (roughly equivalent to $5 USD) for a session lasting approximately 30 min. Pre-screening criteria restricted participation to UK residents, over 18 years of age, with self-reported normal or corrected-to-normal vision. The sample comprised 151 females (73%) and 55 males (26%), with a mean age of 31 (*SD* = 12, *Min* = 18, *Max* = 70). Of those 137 (66%) had a full driving licence at the time of the study. Among those who had a licence, the average number of years of driving experience was 15 (*SD* = 12, *Min* = 1, *Max* = 50). Their average annual mileage was 7,190 (*SD* = 6,353).

2.1.2. Design

The study adopted a mixed design: 2 (Operator-between participants) X 3 (Outcome Severity-within participants) X 6 (Scenario-within participants). Each participant was shown six scenarios featuring a range of different hazardous situations in random order.

Outcome Severity was a repeated-measure variable with three levels: Near-miss (the vehicle narrowly avoided a collision), Minor Accident (a collision occurs but no injury caused) and Major Accident (the collision results in injuries). All participants then studied 18 vignettes. Each scenario was presented in a block of three, each one with a different outcome severity. The order of the blocks was randomised, and as was the order of outcome severity within blocks.

Operator was manipulated between-participant by changing the details of the background information of the scenarios: Half of the participants were told that the target vehicle in all scenarios was an "advanced autonomous computer system with an impeccable safety record" (AS Condition), the other half of the participants were told that "the vehicle was driven by an experienced driver also with an impeccable safety record" (HD Condition).

There were three main dependent variables: The blame on the operator of the vehicle, the blame on third parties involved and the degree of trust. In addition to these incident-specific variables, at the outset of each session participants' pre-existing attitudes towards autonomous vehicles (AS condition) or human driver (HD condition) were measured, including the likelihood of use and trust.

2.1.3. Materials

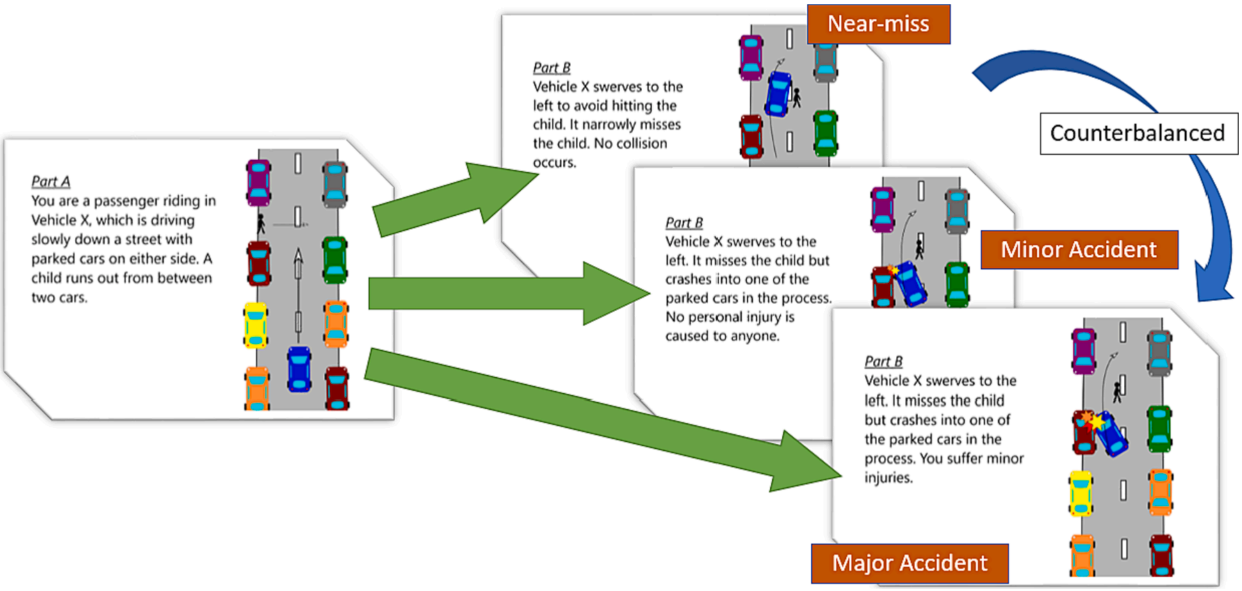
The six scenarios were depicted in the form of a vignette comprising a graphical illustration accompanied by descriptive text (Fig. 1). They were of roughly equivalent complexity and length and they were primarily derived from previous studies (e.g., the Pedestrian scenario), real-life accidents (the Bus scenario) and the hazard perception training (e.g., the Child scenario). Each scenario comprised two parts: Part A described the antecedent conditions in graphical and written form and Part B described the action of the target vehicle as well as the severity of outcome (See Fig. 2). On different trials, Part A of each scenario was coupled with three alternative versions of Part B, corresponding to the three levels of outcome severity. In the example of Fig. 2 (Scenario: Child), 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." For Part B, the Near-miss version read "Vehicle X swerves to the left to avoid hitting the child. It narrowly misses the child. No collision occurs." The Minor Accident version read: "Vehicle X swerves to the left to avoid hitting the child. It narrowly misses the child but crashes into one of the parked cars in the process. No personal injury is caused to anyone". The Major Accident version read: "Vehicle X swerves to the left to avoid hitting the child. It narrowly misses the child but crashes into one of the parked cars in the process. You suffer (minor) injuries." Each Part B also had a graphical illustration of the outcome.

2.1.4. Procedure

The *Prolific* web-link led participants to an introductory page followed by a consent form. Biographical questions followed (e.g., age, gender) along with driving-experience-related questions (e.g., "Do you currently hold a driving licence?" etc.).

Participants were randomly assigned to one of two Operator conditions: In the AS condition, participants' general attitudes towards AVs were measured by ratings on two 11-point scales: one measuring their likelihood of using an AV when the technology becomes available; and the other measuring their trust in the AV technology. Participants in the HD condition were asked two similar questions about their attitudes towards being driven by a human driver.

Participants were then presented with the vignettes described in the material section and were asked to judge blame on the operator of the target vehicle: "Based on the scenario you just experienced, to what extent do you think the driver of/the autonomous system that controls Vehicle X should be blamed for the incident that just took place?" The blame on third parties were measured in a similar



Part A
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
Vehicle X swerves to the left to avoid hitting the child. It narrowly misses the child. No collision occurs.

Part B
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
Vehicle X swerves to the left. It misses the child but crashes into one of the parked cars in the process. You suffer minor injuries.

Fig. 2. Part A of the "Child" scenario, paired with three alternative versions of Part B (Full coloured version available in the digital copy of this paper).

way using an 11-point scale. Participants were also asked how much they trust the driver/autonomous system that controls the vehicle to operate safely in the future. The order of these was randomised and interspersed with questions to evaluate the general attentiveness of the participants (e.g., “What colour was Vehicle X?”, “In which direction did Vehicle X swerve to avoid hitting the deer?”).

At the end of the session, participants were debriefed about the rationale of the study and given an opportunity to leave comments about any aspect of the study.

2.2. Results

The average correct rate in responses to the attention questions was 94.43% ($SD = 0.05$), which suggests the participants acted diligently when they undertook the experiment proper.

Unless stated otherwise, all scores of the key dependent variables were submitted to a 2 (Operator) × 3 (Outcome Severity) × 6 (Scenario) mixed ANOVA.

2.2.1. Blame on autonomous vehicles versus human drivers

Blame as ascribed to ASs and HDs was not universal – their relative levels were to some degree context-dependent, on both scenario and outcome severity. ASs received more blame than HDs in most scenarios but the difference not only changed in magnitude across scenarios but was reversed in one case (the Bus Scenario). As Fig. 3 shows, in five out of six scenarios blame was greater for ASs than HDs ($F(1, 204) = 7.86$, $p = .006$, $\eta^2 = 0.04$) - an effect reversed in the Bus Scenario ($F(1, 204) = 6.81$, $p = .010$, $\eta^2 = 0.03$), which was confirmed by a significant Scenario – Operator interaction ($F(5, 1020) = 8.28$, $p < .001$, $\eta^2 = 0.04$).

Additionally, as Fig. 4 shows, ASs only received more blame when the incident had tangible consequences (i.e., when the incident was either minor or major but not when a near-miss), as evidenced by a significant Outcome Severity - Operator interaction ($F(2, 408) = 12.72$, $p < .001$, $\eta^2 = 0.06$).

2.2.2. Blame on third parties

Blame on some third parties (e.g., the bus driver in the bus scenario and the management of the national park in both the Deer and Tree scenario) seemed to show a reciprocal pattern to the blame on the target vehicle: it was less strong for the AS than for the HD but this was not true for other third parties (e.g., the pedestrian in the bus scenario) as shown by the interaction of Third Party and Operator ($F(7, 2856 = 2.27$, $p = .027$, $\eta^2 = 0.01$) in an analysis involving 2 (Operator) × 3 (Outcome Severity) × 8 (Third Party) mixed ANOVA.

2.2.3. Trust in autonomous vehicles versus human drivers

Ratings of post-incident trust in the driver/operator of the target vehicle were almost a mirror image of the ratings of blame, pointing to the close causal connection of the two. Participants trusted the AS less than the HD and as with blame, the effect of Operator was found to be modulated by Scenario ($F(5, 1020) = 6.12$, $p < .001$, $\eta^2 = 0.03$) (see Fig. 3). The difference between the HD and the AS was smaller in the Bus Scenarios compared to other scenarios, although it was not reversed—as it was in the ratings of blame—for the same scenario. The effect of Operator on trust was also contingent on Outcome Severity ($F(2, 408) = 6.50$, $p = .002$, $\eta^2 = 0.03$) (Fig. 4), with the magnitude of difference between ASs and HDs being smaller after a near-miss compared to the other two outcomes, a pattern consistent with that of blame.

The complementarity of blame and trust is further illustrated by negative correlational coefficients, which show that in the AS condition blame and post-incident trust had a significant correlation in all three outcome conditions (ranging from $r = -0.31$, $p = .002$ to $r = -0.49$, $p < .001$), although the magnitudes of the comparable correlations seem to be slightly greater in the HD condition (ranging from $r = -0.46$, $p < .001$ to $r = -0.49$, $p < .001$).

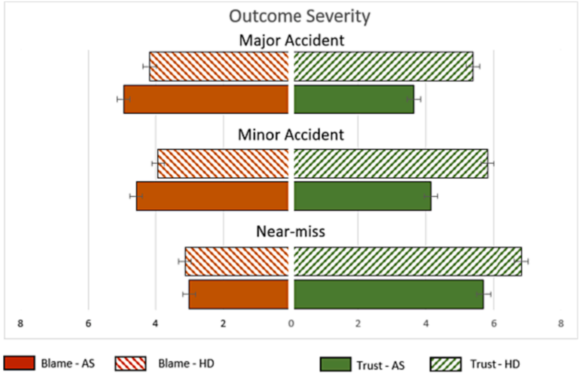


Fig. 3. Mean ratings of blame on the target vehicle and post-incident trust in both the Human Driver and Autonomous System Condition across all scenarios (Error bars = +/- 1 SE; Full coloured version available in the digital copy of this paper).

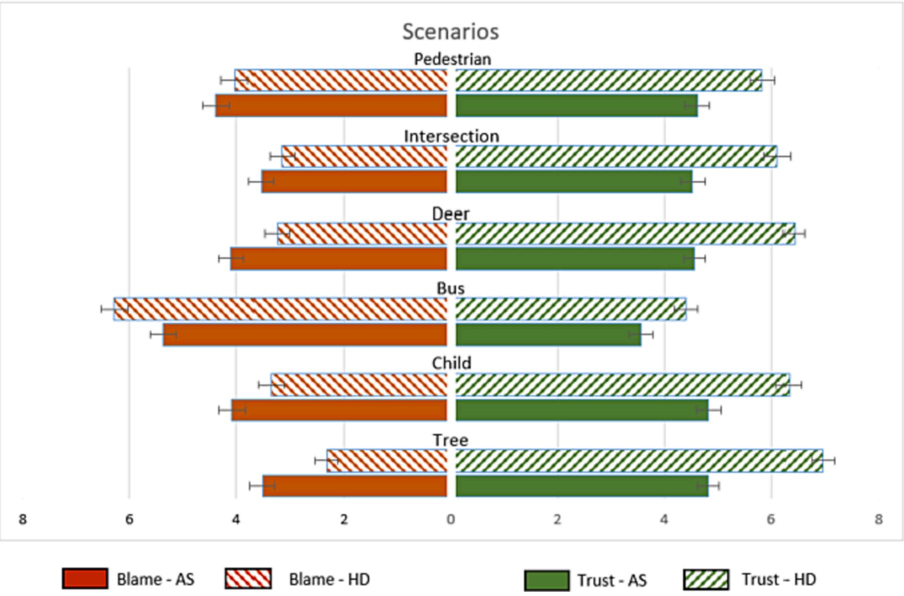


Fig. 4. Mean ratings of blame on the target vehicle and post-incident trust in both the Human Driver and Autonomous System Condition across three levels of outcome severity (Error bars = +/- 1 SE; Full coloured version available in the digital copy of this paper).

2.2.4. *Pre-existing Attitudes' Relation to Blame and Trust*

Blame and trust were determined by different mechanisms, as shown by the correlation coefficients in Table 1. Judgments of blame are not affected by pre-existing attitudes towards ASs or HDs, as they had poor correlations with either the likelihood of using AVs/HDs or the general trust in AVs/HDs. Instead, it may be that blame is dictated primarily by situational factors, that is, on case-by-case inference. This lack of correlation with pre-existing attitudes might in part be the result of the character of the scenarios used in the study, insofar that they were not sufficiently ambiguous with respect to culpability.

In contrast, strong correlations were evident between pre-trial attitudes and trust. This was particularly the case with trust in AVs. There was only the occasional case where pre-trial trust in HDs predicted post-incident trust significantly. Hence, trust in a particular AS is mostly built on trust in the technology in general, whereas for conventional vehicles, trust towards a HD is mostly informed by the behaviour/performance of that individual driver. We base this conclusion on the differences in correlational strength, combined with the finding that blame and post-incident trust correlated more strongly in the human driver condition than in the autonomous system condition. It should be noted that this conclusion is only tentative because the general attitude towards autonomous driving technology was only measured crudely by averaging the responses to two questions (one regarding trust and the other willingness to adopt).

2.3. *Discussion*

The most striking finding of Experiment 1 was that people apply different standards to ASs and HDs when ascribing blame. There was some evidence that this was context dependent, with the bus scenario showing a reversed trend of the other five.

The choices of scenarios in this study did not allow us to engage in strong generalisations about the role of context. Nevertheless, we can speculate that the distinctive feature of the anomalous bus scenario is that it portrayed strong but complex causal antecedents to the accident that should have allowed a vigilant HD to anticipate the accident (Stahl et al., 2014). So, the inference that disembarking bus passengers are likely to cross the road should have served as a strong cue to the HD of the target vehicle of unfolding events—even in the absence of direct visibility of the pedestrians—enough for evasive measures to be taken in a timely way. An accident in these

Table 1
Correlations Between Pre-existing Attitude and the Post-incident Blame and Post-incident Trust in the Autonomous and Human Condition.

N = 103		Aggregated Blame			Aggregated Trust		
		Near-miss	Minor Accident	Major Accident	Near-miss	Minor Accident	Major Accident
Likelihood Of Using	AS	-0.033	0.062	0.130	0.542**	0.572**	0.502**
(Pre-Trial)	HD	0.025	0.123	0.133	-0.014	0.025	-0.027
Trust	AS	-0.046	0.041	0.073	0.603**	0.619**	0.582**
(Pre-Trial)	HD	-0.003	0.149	0.150	0.286**	0.218*	0.117

** significant at $\alpha = 0.01$ level.
* significant at $\alpha = 0.05$ level.

circumstances is especially blameworthy, therefore. The other scenarios in our set had emergencies occurring very suddenly and unpredictably, for which ASs, should be better equipped to avoid, as they depend more on reaction speed and accuracy, rather than the fashioning of inferences.

The bus scenario provides an instance that induces the expectation that humans are expected to better manage causal cues and use them to better anticipate the outcomes of gradually emerging danger. Conversely, ASs are expected to outperform humans when causal cues are weak and hence avoiding an accident depends more on reaction speed and accuracy, rather than inference making. This hypothesis was formally tested in Experiment 2, in which the strength of causal cues was explicitly manipulated.

3. Experiment 2

Experiment 2 extended Experiment 1 by formally testing the moderating effect of causal cue strength on blame and trust. We hypothesized that in incidents preceded by weak causal cues (typically the accident occurring suddenly and without warning), ASs are blamed more (and trusted less) than a HD. This is based on the commonly held supposition that machines and computers outperform humans in the speed and accuracy of their response. When preceded with strong causal cues, in contrast, we expect the contrary: HDs will be assigned more blame (and less trust) because of the expectation that humans are (for now, at least) better at contextualising and anticipating action, particularly if the setting is complex.

3.1. Method

3.1.1. Participants

200 participants were recruited through *Prolific Academic* to take part in the study and paid £7 (GBP) each (roughly equivalent to $10 USD) for a session lasting roughly 60 min. The pre-screening criteria were as in Experiment 1. The sample consisted of 125 females and 74 males (one declined to reveal gender), with a mean age of 34 yrs (SD = 13, Min = 18, Max = 73). The sample included 152 participants with a full driving licence at the time of the study, with an average driving experience of 18 yrs (SD = 15yrs, Min = 1, Max = 60) and an average annual mileage of driving 6,727 (SD = 5,050).

3.1.2. Design

This study adopted a similar design as Experiment 1 but with four scenarios. It introduced a new within-participant independent variable *Causal Cue Strength* with two levels (Strong *vs.* Weak) operationalised via scenario narratives. Levels of outcome severity were reduced to the two conditions (Near-miss versus Major Accident) that showed the largest differences in Experiment 1. Else, the method was as in Experiment 1. Each of the Analyses of Variance, below, contains all four factors.

3.1.3. Materials

In overall format, the vignettes used in this study resembled those in Experiment 1: Part A depicting the emergency situation and Part B the action of the vehicle and outcome. *Causal Cue Strength* was manipulated by creating strong and weak causal narratives in Part A. For example, in the Pedestrian scenario, a pedestrian stepped in front of the target vehicle, in the strong cue version they faced into

Table 2
Textual narratives in the two causal cue strength conditions across four scenarios.

Scenarios	Strong Causal Cues	Weak Causal Cues
Pedestrian	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a pedestrian standing on the edge of the pavement on the **left-hand side** of the road, facing **towards** the road and **looking both ways**. As your vehicle is about to go past, the pedestrian steps into the road in front of your vehicle.	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a pedestrian standing on the edge of the pavement on the **left-hand side** of the road, facing **away from** the road and **looking both ways**. As your vehicle is about to go past, the pedestrian steps backward into the road in front of your vehicle.
Conversation	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle two pedestrians standing on the edge of the pavement on the **left-hand side** of the road, who seem to be having a **fierce quarrel**. As your vehicle is about to go past, one of the them steps into the road in front of your vehicle.	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle two pedestrians standing on the edge of the pavement on the **left-hand side** of the road, who seem to be having a **pleasant conversation**. As your vehicle is about to go past, one of the them steps into the road in front of your vehicle.
Bus/Lorry	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a **bus** stopped on **your side** of the road. As your vehicle goes around the bus, some pedestrians step out into the road from in front of the bus.	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a **lorry** stopped on **your side** of the road. As your vehicle goes around the lorry, some pedestrians step out into the road from in front of the lorry.
Child	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a child standing on the edge of the pavement on the **left-hand side** of the road, facing **towards** the road and **waving a hand**. As your vehicle is about to go past the child, another child runs out from the **right-hand side** into the road in front of your vehicle.	You are a passenger riding in Vehicle X, which is driving down a town road. You spot ahead of your vehicle a child **sitting** on the edge of the pavement on the **left-hand side** of the road, **reading a book**. As your vehicle is about to go past the child, another child runs out from the **right-hand side** into the road in front of your vehicle.

the road and looking both ways before stepping out, whereas in the weak cue version, the pedestrian was facing away from the road when looking both ways. Table 2 gives the narratives associated with the Part A of each scenario. *Type of Operator* and *Outcome Severity* were manipulated in a similar fashion as in Experiment 1. As before, measures were taken of blame and trust.

3.1.4. Procedure

This was as Experiment 1, except that participants initially undertook a practice trial using a scenario unrelated to those used in the experiment proper.

3.2. Results

The average correct rate in responses to the attention questions was 95.69% ($SD = 0.04$), which implies that participant's responses are trustworthy.

For economy as well as clarity of exposition, main effects from analyses of variance are not reported when they are also part of a significant interaction. Additionally, although interactions are found with scenario type and third parties, we do not report these on the grounds that the differences were insufficiently systematic.

3.2.1. The effect of causal cue strength on blame

As in Experiment 1 ASs tended to be blamed more than human drivers when the outcome of the incident was an accident with a tendency for the reverse to be true when the incident was a near-miss (see Fig. 5), which was confirmed by a significant interaction of Operator and Outcome ($F(1, 198) = 19.07$, $p < .001$, $\eta^2 = 0.09$). Blame was higher when the causal cue was strong but the magnitude varied across scenarios, as evidenced by a significant interaction between Causal Cue Strength and Scenario ($F(3, 594) = 10.88$, $p < .001$, $\eta^2 = 0.05$).

Most importantly, when causal cues were weak, the participants assigned more blame to autonomous systems than to human drivers, while the opposite was true when causal cues were strong (see Fig. 6). This was confirmed by a Causal Cue Strength - Operator interaction which was significant on the level of $\alpha = 0.05$ ($F(1, 198) = 4.68$, $p = .032$, $\eta^2 = 0.02$). This pattern is consistent with our hypothesis regarding the modulating effect of causal cue strength on blame attribution.

Mirroring the blame on the target vehicle, the blame on third-parties was generally lower when the causal cues were strong than when they were weak an effect that varied across third-parties ($F(4, 792) = 14.20$, $p < .001$, $\eta^2 = 0.07$). However, unlike results with blame on the target vehicle, the interaction of Causal Cue Strength and Operator was not significant ($F(1, 198) = 0.03$, $p = .870$, $\eta^2 < 0.01$).

3.2.2. The effect of causal cue strength on trust

The effect of Causal Cue Strength has affected trust in a similar fashion to blame. Weak causal cues engendered more trust when the operator was a human driver than when an AV (see Fig. 6). The magnitude of this effect is reduced (although not reversed) when causal cues were strong ($F(1, 198) = 13.87$, $p < .001$, $\eta^2 = 0.07$). This pattern of the interaction is consistent with the idea that stronger causal cues weaken tendency to trust human drivers more than autonomous systems after a traffic incident. The fact that trust didn't display a perfect reciprocal pattern to blame echoes with the findings of Experiment 1 and indicates that trust is not completely informed by blame – other factors, like pre-existing attitudes, might also have played a role.

4. General discussion

The experiments reported here had a two-fold purpose: to establish whether, in making judgments of blame or trust after an accident, different standards are applied to autonomous systems (ASs) and human drivers (HDs); and second, whether the discrimination

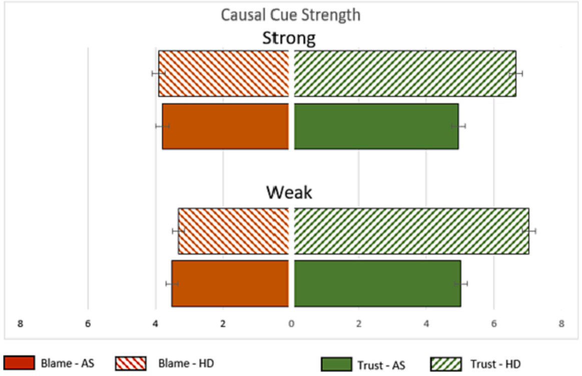


Fig. 5. Mean ratings of blame on the target vehicle and post-incident trust in both the Human Driver and Autonomous System Condition across two levels of outcome severity (Error bars = +/- 1 SE; Full coloured version available in the digital copy of this paper).

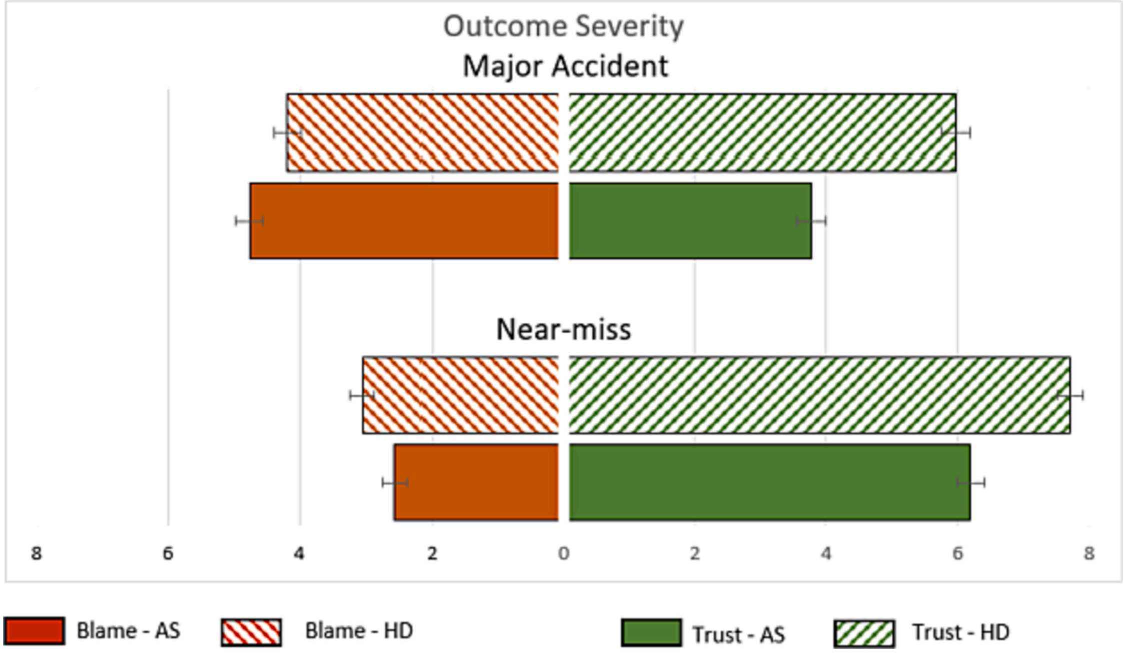


Fig. 6. Mean ratings of blame on the target vehicle and post-incident trust in both the Human Driver and Autonomous System Condition across two levels of causal cue strength (Error bars = +/- 1 SE; Full coloured version available in the digital copy of this paper).

is context-dependent and if so, whether the nature of this dependency conforms with the common perceptions of the strengths and weaknesses of machine versus humans in different scenarios (e.g., of various strength of causal cues). We found supporting evidence for both conjectures.

In Experiment 1, participants exhibited an inclination to assign more blame on autonomous systems in five out of six scenarios: They blamed an autonomous system more than a human driver after a traffic incident even though in all respects the antecedent events were identical. This effect became more pronounced as the outcomes of the incident were more consequential for the pedestrian. Experiment 1 contradicted previous findings (Franklin et al., 2021; Hong, 2020) by demonstrating that the asymmetry of blame is not uni-directional. There was a significant reversal in the Bus scenario: autonomous systems were judged more favourably than human drivers. This supported our 'capability hypothesis': that in judging blame participants incorporate estimates of the capability of a given technology, instead of applying a higher judgmental standard against the technology in a universal way.

This capability hypothesis was tested in Experiment 2 by varying the strength with which cues in the narrative helped presage the outcome. When the causal cue was strong, the participants tended to blame the human driver more than the autonomous system but the reverse was true when the causal cue was weak. While autonomous vehicles are expected to surpass humans in the speed and accuracy of their actions, humans are expected to make better judgments, by drawing on human associative reasoning and on prior experience. In turn, this highlights the role of reasoning about the role of capability context rather than the blanket application of a pre-formed prejudice about a technology.

One implication that these results might carry is that making an autonomous system more human-like might change the expectations that observers have of it and hence change the assignment of blame. Such an anthropomorphic approach could be implemented by augmenting the autonomous system with a human-like voice, or speech-recognition capability, or indeed humanoid robots that combine these capacities with gesture and facial expression. Whether such changes constitute a misrepresentation of the system's true character and is thus deceitful, is perhaps moot.

A secondary objective was the study of how blame attribution colours trust in autonomous systems more generally. We reasoned that greater blame should translate into more severe damage to post-incident trust. This idea received qualified support from both experiments inasmuch as there was a reciprocal relation of ratings of blame to those of trust. Perhaps the contrived setting of an experiment and the relatively low fidelity of the descriptions militated against expression of long-term consequences for trust. The relationship found here makes us optimistic that when we go onto scenarios using high-quality graphic simulations of accidents the association of blame and trust will become stronger.

Pre-existing attitudes towards autonomous vehicles or human drivers helped to shape trust. Interestingly, these correlations were much stronger in the autonomous system group than in the human driver group. Correlations between blame and trust were weaker in the autonomous system condition compared to the human driver condition. Together, these findings suggest that trust in an autonomous system is built on trust of autonomous technology in general, whereas for conventional vehicles, human trust towards a particular driver is mostly informed by the behaviour/performance of that driver. This difference in the judgmental basis for trust might be the manifestation of a more fundamental difference in human perception between autonomous vehicles and conventional vehicles: while people might perceive that there is a great diversity in driving skills among human drivers ('There are good drivers and not so good drivers'), perceptions of the operational competence of autonomous vehicles appears to be more homogenous ('All autonomous vehicles are dangerous.'). This perception might be the descendant of the human stereotype of mass-produced industrial products, where efficiency and profitability are achieved through standardisation and homogeneity. In the world of conventional automobiles, design or production defects often affect more than one unit and collective recalls from manufacturers are not uncommon. It could be that this stereotype has been transferred from the perception of conventional vehicles to that of autonomous vehicles, even though such transfer is not always warranted - The machine-learning algorithms underpinning the control systems of autonomous vehicles might render each vehicle different from one another depending on what data has been used in the training and where the vehicle has been used and for what purpose(s).

This divergence in the judgment of trust is also analogous to a social psychological construct known as the 'out-group homogeneity' (Park & Rothbart, 1982; Rubin & Badea, 2012), namely that humans tend to perceive that members of their own social group (the 'in-group') to be more diverse and differentiated than a social group with whom they less easily identify (the 'out-group'). Analogously, the participants might have viewed human drivers as an "in-group" and autonomous systems as an "out-group" and hence are more readily to assume homogeneity for the latter ('Autonomous vehicles are all the same'). Machines of the same class are more homogenous than animals of the same class, such as humans.

Regardless of the causes, this perception of homogeneity of autonomous vehicles will likely to increases the tendency to generalise the attributes of one unit to the whole group (Quattrone & Jones, 1980). Whereas an encounter with one delinquent human driver is unlikely to affect attitudes towards all drivers, human perception of autonomous vehicles as well as the AI technology in general is more likely to be influenced by the isolated case of one 'bad' autonomous system. While suggestive, follow up work should examine this notion more systematically. For example, quantitative models should be built and used to formally investigate the relationship between the appraisal of the driving behaviour of one specific autonomous vehicle and general attitudes towards the technology. Such investigations should employ more comprehensive and sophisticated measures of the subjective experience of individual autonomous vehicles - e.g., Checklist for Trust (Jian et al., 2000) – as well as those to measure general attitudes – e.g., Autonomous Vehicle Acceptability Scale (AVAS) (Qu et al., 2019)- which would - arguably - reveal a more nuanced picture of how individual factors of the perception of one autonomous vehicle (e.g., perceived riskiness/usefulness) inform the perception of the technology as a whole.

At the time of writing the use of autonomous vehicles is not widespread, but media reports of the few accidents with prototype systems (e.g., Tesla electric SUV crash on US Highway 101, Mountain View, California, 2018) have usually failed to highlight their rarity and the likelihood of fewer road traffic accidents as autonomous vehicles become more widely adopted. Negative perceptions of the technology are likely to diminish its adoption, a tendency strengthened in the case of autonomous vehicles by direct experience or hearsay accounts of accidents. Paradoxically perhaps, this effect of accidents on autonomous vehicle adoption (as well as kindred effects on their continued use) will become more impactful because of their low incidence. It is a well-documented phenomenon in the literature of cognitive heuristics and biases that people's subjective estimation of the likelihood of an adverse outcome (e.g., autonomous vehicles having an accident) can be overly inflated if examples of such could be easily brought to mind (Kahneman & Frederick, 2002; Tversky & Kahneman, 1973). Hence, the low frequency of autonomous vehicle accidents may actually make them more accessible in memory and inflate their subjective likelihood. Human risk perception could also be distorted by the potential negative emotional impact of an adverse event (Slovic et al., 2004). For example, people may over-react to "dread risk" (Gigerenzer, 2004)– adverse events that have low probably but severe consequences (e.g., plane crashes) due to the negative emotions they provoke. The detailed, and often graphic media coverages of traffic incidents involving autonomous vehicle are likely to induce dread, fear and anxiety that exacerbate avoidance behaviours despite measured judgments of accumulated and statistically sound data indicating the contrary.

The present findings bear important implications for legislation, policy making and the design principles of autonomous systems. The results from the juxtapositions of human drivers and autonomous systems in different situations provide useful insights into how the current legal systems, which are based on the assumption of human mental/physical capacity and limitations, need to be adjusted to accommodate the advent of fully autonomous vehicles. The product liability laws and tort laws in many countries (e.g., UK) prescribe a consideration of "reasonable expectation" before a product (e.g., a vehicle) can be claimed defective (Gurney, 2013; Kysar, 2003). For example, Section 3 of the Consumer Protection Act 1987 in the UK provides that a product is deemed to be defective when "the safety of the product is not such as persons generally are entitled to expect". However, "reasonableness" and "entitlement" are vague terms and they hinge largely on experiences and intuitions. What should be the reasonable braking distance expected of the vehicle given the existing technologies? What should be the reasonable computing speed of the decision system of an autonomous vehicle and sensitivity of its sensors? The findings from our studies provide sharp contrast between human drivers and autonomous systems with respect to the public perceptions of their capabilities. Perhaps it is not surprising to find the general trend from our results that the public expect superior safety performance from autonomous systems than human drivers; after all, this is one of the reasons why they were conceived. However, in many situations autonomous systems might be expected to underperform human drivers, in which people might react more leniently to the mistakes of autonomous vehicles. These expectations and perceptions might be naïve and biased but will have tangible impacts on the consequences of lawsuits as well as the long-term acceptance and continuous uptake of autonomous vehicles.

Our findings also point to the necessity of updating the public knowledge and dispersing misconceptions. Capabilities of autonomous vehicles change apace, a dynamism that presents considerable challenges to policy makers and regulatory organisations. Laws and public perceptions will need to be calibrated all the time through campaigns and educational programmes. One important responsibility must rest on how autonomous vehicle manufacturers/designers manage consumer expectations by providing an accurate portrayal of the true capabilities of their products. This point is well illustrated by the several traffic accidents involving TESLA semi-autonomous cars as the result of user misuse of their automated functions. Although the fact that the car needs constant human monitoring was explicitly written into the user manual, the name “Autopilot”, which TESLA has given to it, may have caused confusion and over-reliance on this feature. Thus, technical limitations of autonomous vehicles and any other autonomous systems should be saliently communicated to their users in a timely fashion through the design of the product.

5. Conclusion

We present findings of two experiments demonstrating that 1) people adopt different standards when making judgments of blame on autonomous systems as opposed to human drivers in the event of road accidents; and 2) the standards being adopted are a reflection of the perceived capacity of autonomous systems and of human drivers, which are applied in a context-specific way. When human drivers fail to use their capacity for anticipation and extrapolation, they are more severely blamed than an autonomous system in the same setting. When speed and accuracy is at a premium—coupled with the expectation that these are activities at which machines excel—autonomous systems receive more blame for an accident. Judgment of trust regarding a specific autonomous system after a traffic incident is informed by the judgment of blame. But unlike trust in human drivers, trust in an autonomous system is also largely shaped by the trust in the autonomous technology in general. This might be caused by the fact that in terms of the variation of performance, machines are perceived as a homogenous set (and humans as a heterogenous one). This perception could also lead to the belief that single instances of the (bad/negative) behaviour of machines are more representative of the future behaviour of the class as a whole. Hence, autonomous vehicle accidents will likely contribute more to the erosion of trust in subsequent encounters, than those involving human drivers –and it is key that we develop methods to alleviate this potential problem given the many positive factors that could be realised by the use of autonomous vehicles in the future. The low frequency and the intensive media coverage of these incidents could distort the public’s risk perception of the technology, which highlights the importance of educating the public regarding the autonomous vehicles’ capabilities in a timely fashion – with research like that reported within the current paper offering an important contribution to a growing evidence base that also has great potential to inform policy, practice and many other factors relating to the design, deployment, regulation, use and further development of such technologies.

CRediT authorship contribution statement

Qiyuan Zhang Conceptualization, Data Curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft, Writing – review & editing. **Chris D. Wallbridge** Conceptualization, Investigation, Methodology, Visualization, Writing – original draft. **Dylan M. Jones** Conceptualization, Funding Acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Phillip L Morgan** Conceptualization, Funding Acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Abe, G., Sato, K., & Itoh, M. (2015). Driver’s trust in automted driving when passing other traffic objects. *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 897–902. https://doi.org/10.1109/SMC.2015.165.

Adnan, N., Md Nordin, S., bin Bahruddin, M. A., & Ali, M. (2018). How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. Transport. Res. Part A: Policy Pract. 118, 819–836. https://doi.org/10.1016/J.TRA.2018.10.019.

Anderson, J.M., Nidhi, K., Stanley, K.D., Oluwatola, O.A., Samaras, C., Sorensen, P., 2014. Autonomous vehicle technology: A guide for policymakers. RAND Corporation. https://doi.org/10.7249/j.ctt5hhwgz.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.F., Rahwan, I., 2018. The moral machine experiment. Nature 563 (7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J.B., Shariff, A., Bonnefon, J.F., Rahwan, I., 2020. Drivers are blamed more than their automated cars when both make mistakes. Nat. Hum. Behav. 4 (2), 134–143. https://doi.org/10.1038/s41562-019-0762-8.

Bellet, T., Cunneen, M., Mullins, M., Murphy, F., Pütz, F., Spickermann, F., Braendle, C., Baumann, M.F., 2019. From semi to fully autonomous vehicles: New emerging risks and ethico-legal challenges for human-machine interactions. Transport. Res. F: Traffic Psychol. Behav. 63, 153–164. https://doi.org/10.1016/j.trf.2019.04.004.

Bennett, J.M., Challinor, K.L., Modesto, O., 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., & Rahwan, I. (2019). The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. *Proceedings of the IEEE*, *107*(3), 502–504. https://doi.org/10.1109/JPROC.2019.2897447.

Bonnefon, J.F., Shariff, A., Rahwan, I., 2016. The social dilemma of autonomous vehicles. Science 352 (6293), 1573–1576. https://doi.org/10.1126/science.aaf2654.

Choi, J.K., Ji, Y.G., 2015. Investigating the importance of trust on adopting an autonomous vehicle. Int. J. Hum. Comput. Interact. 31 (10), 692–702. https://doi.org/10.1080/10447318.2015.1070549.

D’Olimpio, L., 2018. Trust as a virtue in education. Educ. Philos. Theory 50 (2), 193–202. https://doi.org/10.1080/00131857.2016.1194737.

de Visser, E.J., Pak, R., Shaw, T.H., 2018. From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. Ergonomics 61 (10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725.

de Winter, J.C.F., Hancock, P.A., 2015. Reflections on the 1951 Fitts list: Do humans believe now that machines surpass them? Procedia Manuf. 3, 5334–5341. https://doi.org/10.1016/j.promfg.2015.07.641.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv Preprint*.

Elish, M.C., 2019. Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). SSRN Electron. J. https://doi.org/10.2139/SSRN.2757236.

Eriksson, A., Stanton, N.A., 2017. Takeover time in highly automated vehicles: Noncritical transitions to and from manual control. Hum. Factors 59 (4), 689–705. https://doi.org/10.1177/0018720816685832.

Fiske, S.T., Taylor, S.E., 1991. Social cognition. Mcgraw-Hill Book Company.

Fitts, P.M., Viteles, M.S., Barr, N.L., Brimhall, D.R., Finch, G., Gardner, E., Grether, W.F., Kellum, W.E., Stevens, S.S., 1951. Human engineering for an effective air-navigation and traffic-control system. In: Fitts, P.M. (Ed.), Human Engineering for an Effective Air-navigation and Traffic-control System. National Research Council, Div. of.

Forster, Y., Naujoks, F., Neukum, A., 2017. Increasing anthropomorphism and trust in automated driving functions by adding speech output. IEEE Intelligent Vehicles Symposium, Proceedings 365–372. https://doi.org/10.1109/IVS.2017.7995746.

Franklin, M., Awad, E., Lagnado, D., 2021. Blaming automated vehicles in difficult situations. IScience 24 (4), 102252. https://doi.org/10.1016/J.ISCI.2021.102252.

Furlough, C., Stokes, T., Gillan, D.J., 2019. Attributing blame to robots: I. the influence of robot autonomy. Hum. Factors 63 (4), 592–602. https://doi.org/10.1177/0018720819880641.

Geisslinger, M., Poszler, F., Betz, J., Lütge, C., Lienkamp, M., 2021. Autonomous driving ethics: From trolley problem to ethics of risk. Philosophy & Technology 2021, 1–23. https://doi.org/10.1007/S13347-021-00449-4.

Gigerenzer, G., 2004. Dread Risk, September 11, and Fatal Traffic Accidents. Psychol. Sci. 15 (4), 286–287. https://doi.org/10.1111/j.0956-7976.2004.00668.x.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80–89. https://doi.org/10.1109/DSAA.2018.00018.

Gkartzonikas, C., Gkritza, K., 2019. What have we learned? A review of stated preference and choice studies on autonomous vehicles. Transportation Research Part C: Emerging Technologies 98, 323–337. https://doi.org/10.1016/j.trc.2018.12.003.

Gold, C., Körber, M., Hohenberger, C., Lechner, D., Bengler, K., 2015. Trust in automation – Before and after the experience of take-over scenarios in a highly automated vehicle. Procedia Manuf. 3, 3025–3032. https://doi.org/10.1016/j.promfg.2015.07.847.

Gurney, J. K. (2013). Sue my car not me: Products liability and accidents involving autonomous vehicles. In *Journal of Law, Technology and Policy* (Vol. 2013, Issue 2).

Hancock, P.A., 2019. Some pitfalls in the promises of automated and autonomous vehicles. Ergonomics 62 (4), 479–495. https://doi.org/10.1080/00140139.2018.1498136.

Hancock, P. A., Nourbakhsh, I., & Stewart, J. (2019). On the future of transportation in an era of automated and autonomous vehicles. *Proceedings of the National Academy of Sciences - PNAS*, *116*(16), 7684–7691. https://doi.org/10.1073/pnas.1805770115.

Hartwich, F., Beggiato, M., Krems, J.F., 2018. Driving comfort, enjoyment and acceptance of automated driving – Effects of drivers' age and driving style familiarity. Https://Doi.Org/10.1080/00140139.2018.1441448 61 (8), 1017–1032. https://doi.org/10.1080/00140139.2018.1441448.

Hong, J.W., 2020. Why is artificial intelligence blamed more? Analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. International Journal of Human-Computer Interaction 36 (18), 1768–1774. https://doi.org/10.1080/10447318.2020.1785693.

Hornborg, A., 2021. Objects Don't Have Desires: Toward an Anthropology of Technology beyond Anthropomorphism. Am. Anthropol. 123 (4), 753–766. https://doi.org/10.1111/aman.13628.

Ilková, V., & Ilka, A. (2017). Legal aspects of autonomous vehicles — An overview. *2017 21st International Conference on Process Control (PC)*, 428–433. https://doi.org/10.1109/PC.2017.7976252.

Jian, J.-Y., Bisantz, A.M., Drury, C.G., 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. Int. J. Cogn. Ergon. 4 (1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04.

Kahneman, D., Frederick, S., 2002. Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics and Biases 49–81. https://doi.org/10.1017/CBO9780511808098.004.

Kahneman, D., Miller, D.T., 1986. Norm theory: Comparing reality to its alternatives. Psychol. Rev. 93 (2), 136–153. https://doi.org/10.1037/0033-295X.93.2.136.

Kallioinen, N., Pershina, M., Zeiser, J., Nosrat Nezami, F., Pipa, G., Stephan, A., König, P., 2019. Moral judgements on the actions of self-driving cars and human drivers in dilemma situations from different perspectives. Front. Psychol. 10 (November), 1–15. https://doi.org/10.3389/fpsyg.2019.02415.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, 80–85. https://doi.org/10.1109/ROMAN.2006.314398.

Kim, P.H., Dirks, K.T., Cooper, C.D., 2009. The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. Acad. Manag. Rev. 34 (3), 401–422. https://doi.org/10.5465/AMR.2009.40631887.

Księżak, P., Wojtczak, S., 2022. A Human Being Must Obey the Commands of a Robot! CAVs, Asimov's Second Law and the New Ground-Breaking Ethics. In: Guarda, T., Portela, F., Augusto, M.F. (Eds.), Advanced Research in Technologies, Information, Innovation and Sustainability. Springer Nature Switzerland, pp. 380–393.

Kysar, D.A., 2003. The expectations of consumers. Columbia Law Rev. 103 (7), 1700–1790. https://doi.org/10.2307/3593402.

Latour, B., 1993. We have never been modern. Harvard University Press.

Latour, B., 1996. Aramis, or the Love of Technology. Harvard University Press.

Lee, J., Abe, G., Sato, K., Itoh, M., 2021. Developing human-machine trust: Impacts of prior instruction and automation failure on driver trust in partially automated vehicles. Transport. Res. F: Traffic Psychol. Behav. 81, 384–395. https://doi.org/10.1016/j.trf.2021.06.013.

Lee, J.D., See, K.A., 2004. Trust in automation: Designing for appropriate reliance. In *Human Factors*. https://doi.org/10.1518/hfes.46.1.50_30392.

Liu, P., Du, Y., 2021. Blame attribution asymmetry in human–automation cooperation. Risk Anal. https://doi.org/10.1111/RISA.13674.

Lorenz, L., Kerschbaum, P., Schumann, J., 2014. Designing take over scenarios for automated driving: How does augmented reality support the driver to get back into the loop? Proceedings of the Human Factors and Ergonomics Society Annual Meeting 58 (1), 1681–1685. https://doi.org/10.1177/1541931214581351.

Ly, A.O., Akhloufi, M., 2021. Learning to drive by imitation: An overview of deep behavior cloning methods. IEEE Trans. Intell. Veh. 6 (2), 195–209. https://doi.org/10.1109/TIV.2020.3002505.

MacIntyre, A., 2013. After virtue. A&C Black.

Madhavan, P., Wiegmann, D.A., Lacson, F.C., 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. Hum. Factors 48 (2), 241–256. https://doi.org/10.1518/001872006777724408.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In *Social Thinking and Interpersonal Behavior* (pp. 313–331). https://doi.org/10.4324/9780203139677.

Malle, B.F., Guglielmo, S., Monroe, A.E., 2014. A theory of blame. Psychol. Inq. 25 (2), 147–186. https://doi.org/10.1080/1047840X.2014.877340.

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An Integrative Model of Organizational Trust. Acad. Manag. Rev. 20 (3), 709–734. https://doi.org/10.2307/258792.

McManus, R.M., Rutchick, A.M., 2019. Autonomous vehicles and the attribution of moral responsibility. Soc. Psychol. Personal. Sci. 10 (3), 345–352. https://doi.org/10.1177/1948550618755875.

Merat, N., & Jamson, A. H. (2009). How do drivers behave in a highly automated car? *PROCEEDINGS of the Fifth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, *January 2009*, 514–521. https://doi.org/10.17077/drivingassessment.1365.

Merat, N., Jamson, A.H., Lai, F.C.H., Daly, M., Carsten, O.M.J., 2014. Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. Transport. Res. F: Traffic Psychol. Behav. 27 (PB), 274–282. https://doi.org/10.1016/j.trf.2014.09.005.

Morgan, P.L., Alford, C., Williams, C., Parkhurst, G., Pipe, T., 2018. Manual takeover and handover of a simulated fully autonomous vehicle within urban and extra-urban settings. Adv. Intell. Syst. Comput. 597, 760–771. https://doi.org/10.1007/978-3-319-60441-1_73.

Morgan, R., Hunt, S., 1994. The Commitment-Trust Theory of Relationship Marketing. J. Mark. 58, 20–38. https://doi.org/10.2307/1252308.

Morgan, P.L., Williams, C., Flower, J., Alford, C., Parkin, J., 2019. Trust in an autonomously driven simulator and vehicle performing maneuvers at a T-junction with and without other vehicles. Adv. Intell. Syst. Comput. 786, 363–375. https://doi.org/10.1007/978-3-319-93885-1_33.

Muir, B.M., 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. Ergonomics 37 (11), 1905–1922. https://doi.org/10.1080/00140139408964957.

Nass, C., Steuer, J., Tauber, E.R., 1994. Computers are social actors. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 72–78. https://doi.org/10.1145/191666.191703.

Nastjuk, I., Herrenkind, B., Marrone, M., Brendel, A.B., Kolbe, L.M., 2020. What drives the acceptance of autonomous driving? An investigation of acceptance factors from an end-user's perspective. Technol. Forecast. Soc. Chang. 161 (July), 120319 https://doi.org/10.1016/j.techfore.2020.120319.

Nordhoff, S., Kyriakidis, M., van Arem, B., Happee, R., 2019. A multi-level model on automated vehicle acceptance (MAVA): A review-based study. Theor. Issues Ergon. Sci. 20 (6), 682–710. https://doi.org/10.1080/1463922X.2019.1621406.

Nyhan, R., 2000. Changing the Paradigm: Trust and Its Role in Public Sector Organizations. American Review of Public Administration - AMER REV PUBLIC ADM 30, 87–109. https://doi.org/10.1177/02750740022064560.

Panagiotopoulos, I., Dimitrakopoulos, G., 2018. An empirical investigation on consumers' intentions towards autonomous driving. Transport. Res. Part C: Emerg. Technol. 95, 773–784. https://doi.org/10.1016/j.trc.2018.08.013.

Parasuraman, R., Riley, V., 1997. Humans and Automation: Use, Misuse, DisuseAbuse. Hum. Fact. 39 (2), 230–253. https://doi.org/10.1518/001872097778543886.

Paret, D., Rebaine, H., & Engel, B. A. (2022). Aspects Relating to Autonomous and Connected Vehicles. In *Autonomous and Connected Vehicles: Network Architectures from Legacy Networks to Automotive Ethernet* (pp. 23–80). Wiley. https://doi.org/10.1002/9781119816140.ch2.

Park, B., Rothbart, M., 1982. Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. J. Pers. Soc. Psychol. 42 (6), 1051–1068. https://doi.org/10.1037/0022-3514.42.6.1051.

Pattinson, J.-A., Chen, H., Basu, S., 2020. Legal issues in automated vehicles: critically considering the potential role of consent and interactive digital interfaces. Humanit. Soc. Sci. Commun. 7 (1), 153. https://doi.org/10.1057/s41599-020-00644-2.

Penmetsa, P., Adanu, E.K., Wood, D., Wang, T., Jones, S.L., 2019. Perceptions and expectations of autonomous vehicles – A snapshot of vulnerable road user opinion. Technol. Forecast. Soc. Chang. 143, 9–13. https://doi.org/10.1016/j.techfore.2019.02.010.

Pöllänen, E., Read, G.J.M., Lane, B.R., Thompson, J., Salmon, P.M., 2020. Who is to blame for crashes involving autonomous vehicles? Exploring blame attribution across the road transport system. Ergonomics 63 (5), 525–537. https://doi.org/10.1080/00140139.2020.1744064.

Qu, W., Xu, J., Ge, Y., Sun, X., Zhang, K., 2019. Development and validation of a questionnaire to assess public receptivity toward autonomous vehicles and its relation with the traffic safety climate in China. Accid. Anal. Prev. 128, 78–86. https://doi.org/10.1016/j.aap.2019.04.006.

Quattrone, G.A., Jones, E.E., 1980. The perception of variability within in-groups and out-groups: Implications for the law of small numbers. J. Pers. Soc. Psychol. 38 (1), 141–152. https://doi.org/10.1037//0022-3514.38.1.141.

Robbennolt, J.K., 2000. Outcome severity and judgments of "responsibility": A meta-analytic review. J. Appl. Soc. Psychol. 30 (12), 2575–2609. https://doi.org/10.1111/J.1559-1816.2000.TB02451.X.

Roese, N.J., Olson, J.M., 1997. Counterfactual thinking: The intersection of affect and function. Adv. Exp. Soc. Psychol. 29 (C), 1–59. https://doi.org/10.1016/S0065-2601(08)60015-5.

Rubin, M., Badea, C., 2012. They're all the same!. but for several different reasons: A review of the multicausal nature of perceived group variability. Current Directions in Psychological Science : A Journal of the American Psychological Society 21 (6), 367–372. https://doi.org/10.1177/0963721412457363.

SAE International. (2018). Taxonomy and definitions for terms related to driving automation systems for on-Road motor vehicles. In *SAE International*.

Sanna, L.J., Turley, K.J., 1996. Antecedents to spontaneous counterfactual thinking: Effects of expectancy violation and outcome valence. Pers. Soc. Psychol. Bull. 22 (9), 906–919. https://doi.org/10.1177/0146167296229005.

Scanlon, T., 2008. Moral dimensions : Permissibility, meaning, blame. Harvard University Press.

Schaefer, K. E., & Straub, E. R. (2016). Will passengers trust driverless vehicles? Removing the steering wheel and pedals. *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 159–165. https://doi.org/10.1109/COGSIMA.2016.7497804.

Schoettle, B., & Sivak, M. (2014). A survey of public opinion about autonomous and self-driving vehicles in the US, UK and Australia. *UMTRI, Transportation Research Institute*, *July*, 1–38.

Sheridan, T.B., 2019. Extending Three Existing Models to Analysis of Trust in Automation: Signal Detection, Statistical Parameter Estimation, and Model-Based Control. Hum. Factors 61 (7), 1162–1170. https://doi.org/10.1177/0018720819829951.

Shionoya, Y., 2001. Trust as a Virtue. In: Shionoya, Y., Yagi, K. (Eds.), Competition, Trust, and Cooperation. Springer, Berlin Heidelberg, pp. 3–19.

Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G., 2004. Risk as Analysis and Risk as Feelings: Some Thoughts about Affect, Reason, Risk, and Rationality. Risk Anal. 24 (2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x.

Stahl, P., Donmez, B., Jamieson, G.A., 2014. Anticipation in driving: The role of experience in the efficacy of pre-event conflict cues. IEEE Trans. Hum.-Mach. Syst. 44 (5), 603–613. https://doi.org/10.1109/THMS.2014.2325558.

Tversky, A., Kahneman, D., 1973. Availability: A heuristic for judging frequency and probability. Cogn. Psychol. 5 (2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9.

Waytz, A., Heafner, J., Epley, N., 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. J. Exp. Soc. Psychol. 52, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005.

Xu, Z., Zhang, K., Min, H., Wang, Z., Zhao, X., Liu, P., 2018. What drives people to accept automated vehicles? Findings from a field experiment. Transport. Res. Part C: Emerg. Technol. 95, 320–334. https://doi.org/10.1016/J.TRC.2018.07.024.

Zablocki, É., Ben-Younes, H., Pérez, P., & Cord, M. (2021). Explainability of vision-based autonomous driving systems: Review and challenges. *ArXiv Preprint*.