Transportation Research Part A 205 (2026) 104885

Contents lists available at ScienceDirect

Transportation Research Part A

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

Should AI be allowed to break the law? Public acceptance and tolerance of discretionary actions performed by autonomous vehicles in response to situations varying by degree of acuteness

Qiyuan Zhang[a,b,c,d], Victoria Marcinkiewicz[a,b,c], Louise Bowen[a,c], Theodor Kozlowski[a,b,c], Tatsuhiko Inatani[e], Yoshiyuki Ueda[f], Hirofumi Katsuno[g], Minoru Asada[h], Phillip L. Morgan[a,b,c,d,i,j,*]

[a] School of Psychology, Cardiff University, 70 Park Place, Cardiff CF10 3AT, UK
[b] Cardiff University Centre for AI, Robotics and Human-Machine Systems (IROHMS), UK
[c] Human Factors Excellence (HuFEx) Research Group, UK
[d] Cardiff University Digital Transformation Innovation Institute (DTII), UK
[e] Graduate School of Law / Faculty of Law, Kyoto University, Japan
[f] Institute for the Future of Human Society, Kyoto University, Japan
[g] Faculty of Social Studies, Department of Media, Journalism and Communications, Doshisha University, Japan
[h] Symbiotic Intelligent System Research Center, Open and Transdisciplinary Research Initiatives, The University of Osaka, Japan
[i] Luleå University of Technology - Psychology, Division of Health, Medicine & Rehabilitation, Regnbagsallen 5, 977 54 Lulea, Sweden
[j] Faculty of Education, Science, Technology and Mathematics, University of Canberra, Australia

ARTICLE INFO

Keywords:
Autonomous driving
Self-driving car
Autonomous system
Discretionary action
Blame attribution
Trust in automation

ABSTRACT

Many countries are on the verge of introducing highly autonomous vehicles (AVs) into current traffic networks dominated by human-driven vehicles. The complex and dynamic nature of road traffic situations requires AVs to exhibit human-like, discretionary behaviours that may fall outside the scope of formal Rules of the Road (e.g., straddling solid lines to let an emergency-services vehicle pass). It is important to understand public attitudes towards these behaviours especially when they may lead to negative outcomes. The current paper presents three experiments in which participants' judgements of blame and trust were probed after being presented with traffic scenarios where AVs or human drivers chose to perform (Experiment 1 & 2) or not perform (Experiment 3) legal or illegal discretionary actions (DAs) to avoid a danger or optimize traffic flow, with various consequences. The results reveal that AVs were blamed more and trusted less than human drivers for performing the same illegal DAs. But with legal DAs, this difference was contingent on the acuteness of traffic situations, hinting toward a shift of judgemental focus from the justifiability of an action to the quality of their execution. Additionally, witnessing AVs performing (or not performing) DAs could potentially improve or worsen their general acceptance depending on the outcomes of the DAs. Our findings paint a promising picture of allowing adaptive behaviours of AVs yet highlight the need to establish formal protocols for designing, regulating, and appraising DAs of AVs as well as the necessity of improving the transparency of their decision-making processes for users.

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

https://doi.org/10.1016/j.tra.2026.104885
Received 9 April 2025; Received in revised form 22 September 2025; Accepted 19 January 2026
Available online 25 January 2026

1. Introduction

Highly autonomous vehicles (AVs) can and will be able to drive themselves with no human intervention under most (Level 4, SAE, 2021) and quite possibly all (Level 5, SAE, 2021) conditions. The wide acceptance, adoption and continued use of such technology hinges heavily on public trust. Despite their many potential benefits, including improving road safety, increasing productivity of the users, enhancing mobility for population samples who cannot drive or find it difficult to drive, alleviating traffic congestion and improving air quality (Ahmed et al., 2022; HM Government, 2022; NHTSA, 2017; Othman, 2022), AVs still face a large degree of public skepticism and resistance, including those who could be potential future adopters. In a recent survey conducted by the Institution of Mechanical Engineers (2023), ~70% of the sample (more than 2000 UK adults) indicated that they would not feel comfortable riding in an AV with no human control at a speed of 70 miles an hour and only one-third of the sample indicated that they would let an autonomous system take over control if they felt tired or unwell. Another study with a larger and more culturally diverse sample (41,932, from 51 countries) revealed that less than half (44.4%) felt that AVs were somewhat or very safe (Moody et al., 2020). In general, attitudes towards AVs, including trust, have been found to be strong predictors of the acceptance (or indeed barriers) of and intention to use (Adnan et al., 2018; Choi & Ji, 2015). A sense of trust in an automated system, cannot only be derived from statistical evidence but also from personal experience of observing and interacting with the system (Hancock et al., 2011; J. D. Lee & See, 2004; Schaefer et al., 2016). It is therefore crucial for highly autonomous AVs to perform and to be perceived to behave in a trustworthy manner when operating across a range of driving situations and scenarios.

Local traffic laws and Rules of the Road (ROTR) could arguably serve as a good anchor point for AVs to follow since these rules are largely designed to ensure traffic safety and efficiency. Many national and international legislative and regulatory bodies, including the United States National Highway Traffic Safety Administration (NHTSA, 2017), the UK Parliament (Automated Vehicles Act, 2024), and the United Nations Economic Commission for Europe (United Nations Economic Commission for Europe, 2020) have set the general requirements that AVs should be engineered and programmed to exhibit behaviours that are compliant with the local traffic rules. However, simply being a safe and lawful road user is far from enough to gain higher public acceptance and fuller integration of AVs into the traffic network (Bin-Nun et al., 2022; Tennant et al., 2021), especially in the case of a hybrid traffic system where AVs are mixed with human-driven vehicles. At least in the early stages of deployment, highly autonomous AVs are currently (on a small scale) and will be for some time operating within the existing road infrastructure which in most countries is designed for human-driven vehicles. Such traffic environments are often complex and fluid, especially in populous cities, in which AVs will need to interact with many other human road users in an adaptive manner including pedestrians, cyclists, motorcyclists and human drivers operating a wide range of vehicles including public transport, emergency services vehicles, delivery vehicles and so on. During these interactions, if AVs exhibit behaviours that violate expectations, norms or conventions, this could induce user (and observer) discomfort, fear, confusion, and frustration, even if the choices made are safe and do not lead to negative outcomes (Hancock, 2019; Hancock et al., 2019).

As an illustration to this point, imagine a hypothetical scenario in which the crew of an ambulance are trying to rush a patient to hospital, needing to negotiate traffic in a heavily congested area. Many human drivers of non-autonomous vehicles, being highly adaptive and capable of exercising discretion in exceptional circumstances, will likely attempt to give way to the ambulance even though it might involve performing a legal but potentially risky manoeuvre (e.g., moving into another lane, crossing the dashed white lines) or even an illegal action such as mounting a kerb or straddling solid-line road markings when there is no oncoming traffic. However, an AV programmed to follow traffic law strictly without adaptive flexibility would not be able to perform any effective action, potentially grinding the traffic to a halt, preventing the emergency vehicle from continuing its journey. There are multiple other examples, including overtaking other road users such as a cyclist to help with traffic flow or swerving to avoid a hazard ahead such as a pedestrian suddenly stepping onto the road. Noting that simply slowing down or stopping (a default of many current AVs) in these situations could in itself lead to a negative outcome – i.e. a build-up of traffic in the case of the cyclist scenario and potentially a collision in the pedestrian scenario. Inability to exercise discretion in these situations will likely erode public trust in AVs, even when their behaviours comply with traffic rules. That is, the 'firm' rules programmed into AVs could, under some circumstances, lead to very negative outcomes. In 2023, a Waymo AV in *Phoenix* refused to follow a police officer's instruction to be diverted to an alternative route due to an incoming parade, possibly because the system could not recognize the hand signals of the police officer, or the manoeuvre might have been perceived or processed as unusual. Both the police officer and passengers within the AV were frustrated and did not know what to do. Traffic was brought to a halt. Incidents like this will likely erode public trust and threaten the acceptance, adoption and continuous use of AVs.

It is therefore imperative to better understand what constitutes good or desirable driving behaviours of AVs – that is, how AVs *should* behave under a number of circumstances – e.g., in response to situations where an evasive (e.g., swerving to avoid colliding with a pedestrian who suddenly steps into the road) or even courteous (e.g., moving into another lane to allow an emergency services vehicle to pass) action could be made, including the decisions they should make and the manner in which they should execute them. For human drivers, guidance for desirable behaviours and actions is usually based on laws, local customs and considerations of common sense in specific situations (Makridis et al., 2023; Schmitt, 2020; Tennant et al., 2021). To gain public trust, it seems that a similar level of adaptability will need to be built into the behavioural specifications of AVs. However, apart from the technical/engineering difficulties associated with programming an adaptive system, behaviours that deviate from, for example, the or Highway Code are inherently risky and there could be negative outcomes. For example, switching lanes to allow an emergency services vehicle to pass having deemed it safe to do so and then subsequently colliding with another vehicle that does something unexpected (e.g., rolling backwards). Such situations will likely attract a higher level of scrutiny due to their perceived extraordinary nature, even though it is common for human drivers to make similar choices and perform discretionary actions. Hence, it is crucial to understand whether AVs should be permitted – and even encouraged through their programming – to exercise discretion, what the boundary conditions are, and what the level of tolerance of the public would be, especially when such actions inadvertently lead to accidents that could not have been foreseen by an AV or a human driver.

The current paper aims to shed light on these questions by probing judgements of blame and trust across three experiments featuring hypothetical scenarios where AVs exercise discretion (or refrain from doing so) leading to various outcomes (e.g., a subsequent collision or a near miss) and comparing responses to the exact same situations, conditions and actions involving human-driven vehicles. The findings have important implications for AV-related policy, regulation and legislation with respect to how AVs should be designed and programmed, and how human users should be prepared for them (e.g., what experience or training is required).

2. Literature review

2.1. Building trust in AVs from experience of them

Research has identified many motivators for the initial adoption of AVs, including sense of enjoyment (Rahman & Thill, 2024), innovativeness (Türkoğlu & Bilici, 2024), sociability (Tennant et al., 2019) and accessibility (Wu et al., 2020). However, successful acceptance, adoption, and continued use of AVs will likely not be achieved unless public trust, at scale, can be realised from the experience of interacting with the technology (Hancock, 2019; Hancock et al., 2019; Xu et al., 2018), not only when it is performing optimally but also in the event of something going wrong. The role of trust has been extensively examined in the context of interpersonal relationships and is deemed to be the glue that binds human society together and makes collaborative economic activities possible (D'Olimpio, 2018; MacIntyre, 2013; Nyhan, 2000; Shionoya, 2001). Mayer et al. (1995, p. 712) defined trust as the *"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"*. Trust is also essential in maintaining relationships between humans and non-human agents – including machines and automation systems, and is a powerful predictor of their adoption and continued use (J. D. Lee & See, 2004; J. Lee & Moray, 1992; Parasuraman et al., 2008; Parasuraman & Riley, 1997a; Sheridan, 2019). This also applies to human interactions with vehicles at different levels of automation. That is, the acceptance, adoption, and continued use of semi- (e.g., SAE Levels 2–3), highly- (SAE Level 4), or even fully- (SAE Level 5) (SAE, 2021) AVs will largely be determined by the extent to which human users are willing to relinquish control to the vehicle while assuming the risk of the vehicle not acting in an expected manner (e.g., AVs not arriving at the destination safely or on time) (Adnan et al., 2018; Choi & Ji, 2015; Olaverri-Monreal, 2020).

Like interpersonal relationships, trust in automation hinges on many factors, including perceived reliability and predictability (Hancock et al., 2011; J. D. Lee & See, 2004; J. Lee & Moray, 1992; Schaefer et al., 2016), which take time to establish, usually through experience with the system (Hoff & Bashir, 2015; Olaverri-Monreal, 2020). It has been estimated through simulating previous collision data that the introduction of AVs could lead to reductions in crash and injury rates by up to 50% (assuming 10% market penetration) or 90% (assuming 90% market penetration), as well as a marked reduction in accidents caused by pedestrians and cyclists (Fagnant & Kockelman, 2015). Yet, AVs are perceived by many as risky and unsafe (Moody et al., 2020; Tennant et al., 2019). The hesitance to potentially use an AV could be partly due to lack of experience or understanding of how the technology works (Schneble & Shaw, 2021), fear induced by the feeling of not being in control (Nastjuk et al., 2020; Nees, 2016), and disproportionate media coverage on incidents and accidents involving AVs, which are mostly rare (Jelinski et al., 2021), despite a growing body of academic literature (e.g., Moody et al., 2020) highlighting significant potential benefits. The effects of these hindrances could be further confounded by human cognitive features that often lead to biased perception of risk.

Human judgement of risk and probability is often not informed by the understanding of base rates and is instead governed by *System II heuristics* or "hot" thinking (Tversky & Kahneman, 1973). For example, research on the *representativeness heuristic* has shown that intuitive perception of the probability of an event is strongly associated with how easily an instance of that event can be brought to mind, either by imagining or retrieving it from memory (Kahneman & Frederick, 2002). This heavy reliance on anecdotal evidence means that people's perception of AV safety could, by-and-large, be deeply impacted by the experience or knowledge of isolated negative incidents involving them (such as those highlighted within so-called negative news stories), despite these incidents being statistically rare (or at least having the potential to be) compared to collision rates associated with non-autonomous vehicles. This cognitive distortion will continue to be marked due to factors such as non-positive media attention and generally negative framing of some collisions or even near misses, especially via social media, which has a general inclination to focus on novel and negative content (Robertson et al., 2023; Watson et al., 2024).

The impact of negative incidents involving AVs on trust is further exacerbated by the fact that people, in general have perhaps unsurprisingly higher expectations of AVs compared to human drivers, with regard to their capabilities (Penmetsa et al., 2019; Schoettle & Sivak, 2014). AVs are hence blamed more than human drivers when they fail to avoid (even if it is not possible to do so) incidents and accidents, even if they perform the same actions under the same circumstances leading to the same consequences (e.g., Hong, 2020 – although for an exception, see evidence from Zhang et al., 2021, 2024). Given that the tolerance of AVs being involved in accidents is much lower than that of human drivers (Liu et al., 2019a,b; Liu & Du, 2021), to be accepted and adopted at scale by the public, AVs need to be safer than human drivers by a considerable margin. There also likely needs to be enough evidence to reinforce this via, for example, media, word-of-mouth, and particularly through direct experience with the technology.

Notably, the source of a negative experience does not need to be an extreme outcome such as a collision resulting in damage, injury or worse. It can also arise from an AV behaving in a manner that provokes discomfort, confusion or frustration, even though those behaviours do not necessarily result in collisions or even near misses. The behaviours of AVs, including every decision they make and how they execute them, need to be fine-tuned to the desirable specifications of all stakeholders (e.g., manufacturers, vehicle users, other road users) so that trust in them can be built over time, sustained and possibly even restored if lost in the event of infrequent incidents. This is likely key to successfully achieving large-scale deployment of AVs with a significant number of people willing to use and interact with them. Crucial to this goal, is to better understand the perceptions of AVs engaging in actions that involve taking an elevated level of risk in response to situations (e.g., negotiating cyclists, responding to an obstruction, allowing an emergency services vehicle to pass, and so on), even when such actions do not always result in positive outcomes (e.g., when a subsequent unpredicted incident occurs). More specifically, we need to better understand trust in and blame on the AV under such circumstances, compared to human drivers performing identical actions with identical outcomes.

2.2. Desirable behaviours of AVs

Maintaining a safe statistical record (e.g., performing without incident) is necessary but not sufficient to warrant trust. The sense of trust or distrust will also be derived from everyday interactions between a user and the technology in mundane as well as more salient situations. To gain public trust, AVs need to behave not only safely but also in a predictable, transparent, and trustworthy manner, with or without passengers. However, formally specifying *desirable behaviours* for AV manufacturers can be challenging. For conventional manually driven vehicles, the responsibility of exhibiting safe and courteous driving behaviour lies with drivers adhering to the law(s). Driver behaviours are highly adaptive to national and (in some cases) local traffic laws, cultural customs, and emerging needs (Tennant et al., 2021). Devising a universal AV standard for behavioural specifications that suits all jurisdictions is not a trivial task (Bin-Nun et al., 2022). While strides have been taken in establishing regulatory global standards for engineering safe AVs (Automated Vehicles Act, 2024; e.g., International Organization for Standardization, 2025); see also Chakraborty et al., 2020 for a review), achieving a uniform specifications for AVs' driving behaviour still proves difficult, due to the variations of legal driving rules as well as cultural customs across different regions (Smith, 2017). Second, and as pointed out by Hancock (2019), due to differences in sensory capability and processing speed, AVs and their human users might have very different views of safe operation (e.g., speed and distance) in any given situation. That is, what is 'thought' by an AV to be a perfectly safe manoeuvre might be perceived as (perhaps very) dangerous or a near-miss situation by a human. This means that behavioural specifications that fully exploit the capability of AVs may not necessarily lead to the highest levels of trust. Instead, AVs might need to "tune down" their behaviours – despite capabilities – to create a better sense of safety for their human users. The question here is therefore not "How *can* an AV drive?" Rather it is "How *should* an AV drive?" (Bellem et al., 2016, 2018)*?* Some early research has investigated factors linked to AV driving styles, including the manner in which an AV executes a typically mundane manoeuvre (e.g., overtaking), and its effect on the comfort and trust of its users. For example, Abe et al. (2015) found that during an overtake manoeuvre, trust was promoted when AVs applied earlier steering manoeuvres and kept wider lateral distances from the object vehicle being overtaken, compared to the passengers' own baseline manual driving where such tolerances were wider. In line with this, Hartwich et al. (2018) found that a driving style that is familiar to the rider/passenger would promotes comfort, enjoyment, and higher acceptance of the AV.

Another branch of literature focuses on AV decision-making: that is, what an AV should do (be programmed to do) in a given situation instead of *how* it does it. For example, researchers have examined moral dilemmas of the "trolley problem" in the context of autonomous driving (e.g., Awad et al., 2018; Bonnefon et al., 2016, 2019). The idea is that advanced sensors and high computing power would afford an AV the ability to "envision" the consequences of alternative course(s) of action in a 'split second' with a high level of certainty, and arguably to be able to make a more 'optimal' decision (e.g., to collide with and potentially kill a pedestrian stepping out in front of the vehicle or take an evasive action to mount the kerb and crash into a wall thus compromise the lives of its passengers and possibly other pedestrians) than a human driver would, or at least more often on average.

The discussion of moral dilemmas can inform what moral values should be instilled (or programmed) into an AVs for it to act appropriately and 'optimally' even in the event of almost certain conditions that will – one way or another – result in a negative outcome including the loss of human life. However, it does not speak to less extreme situations that an average driver needs to deal with almost on a day-to-day basis. For example, deciding whether to overtake a cyclist in order to get to one's destination faster and possibly to help with traffic flow; moving out of the way of an emergency services vehicle that is trying to get through busy traffic; and, driving around a potentially dangerous road condition such as a body of water following heavy rainfall. These are more often than not everyday mundane driving actions for most car drivers and are rarely life-and-death decisions. However, accumulatively, they still make a collective difference to the efficiency of the road traffic system, safety of the drivers / occupants and other road users, and ultimately impact user and other public trust in the systems (Edelmann et al., 2021).

Moreover, the potential consequences of the decisions made by AVs might not be as certain as what is depicted in a trolley problem. In real-life driving situations, all decisions involve a level of uncertainty and risk. It can be argued that AVs are and will be much better equipped than human drivers to envision and permutate how different scenarios can play out with a higher level of certainty. Even so, risk can never be eliminated from the equation, even with the most mundane decisions like overtaking a cyclist, partly due to the fact that the behaviours of other road users cannot be accurately predicted at all times (e.g., the cyclist could decide to speed up and/or veer further into the lane). Therefore, to function optimally and ensure a smooth traffic flow, AVs (or indeed their manufacturers and/or regulators, and so on) will have to assume a certain level of risk with every decision. It is, however, still not well understood how the public would prefer AVs to incorporate the element of risk into their decision-making algorithms. For example, when would the risk become too high to overtake a car in front driving significantly lower than the speed limit? Should the risk tolerance be changed if the passenger of the AV has very good reason for that to happen (e.g., if its passenger needs to be taken to hospital urgently or if the AV needs to take an evasive action because a pedestrian suddenly steps out in front of the vehicle)? These are examples of situations that typical drivers may have encountered numerous times and will have had to make a decision about what to do. ROTR and traffic laws serve as useful anchor points for desirable driving behaviours since they were written to minimise the risk of a collision. However, ad-hoc risk–benefit evaluations are something that human drivers do on a regular basis before executing almost any manoeuvre, including those that are legal and perhaps on occasion some that are not. It is therefore reasonable to specify the behaviours of AVs in a way that incorporates ROTR as well as, moral values and risk–benefit evaluations that are at least akin to what humans would currently do themselves. The following sections discuss the challenges associated with this approach and the necessity of AVs exercising discretion when it comes to the compliance with ROTR as well as the importance of understanding people's tolerance towards AVs making such decisions which could potentially lead to negative outcomes.

2.3. *ROTR as basis for AVs behavioural specifications*

A starting point for specifying desirable driving behaviours for AVs is the legal ROTR − national and local traffic laws expressed in everyday language (e.g., the UK Highway Code). The appeals are clear: the rules are usually written to minimise the risk of collisions and accidents and enforced by national and local authorities. As such, they must be followed by human drivers. Based on a universal understanding and acceptance (at least much of the time) of these by most human drivers, there could be minimum friction costs for introducing AVs to existing traffic networks if they can respect the same rules.

However, instilling ROTR into an AV presents many challenges. From a technical perspective, the ROTR are written in natural languages, and it can be very difficult to transcribe and formalise at least some of the clauses into programming language (Brunello et al., 2019; Chen et al., 2023). They are also not universally consistent. Instead, they are fragmented in the sense that some of the rules in different jurisdictions can differ from each other (despite overlaps). Even if ROTR can be successfully incorporated into the decision-making of AVs, some are inadequate to prescribe behaviours in all eventualities (Prakken, 2017). For example, since 2022 within the UK, drivers are to 'leave at least 1.5 m (5 feet) when overtaking people cycling at speeds of up to 30mph, and [to] give them more space when overtaking at higher speeds' (Highway Code, 2023 − Rule 163). Whilst this represents a significant improvement in terms of cyclist (and driver) safety, some may ask how is a human driver able to quickly determine whether a cyclist will or will not suddenly start to accelerate or decelerate such that the most optimal overtake distance can vary rapidly from moment to moment, not to mention being able to judge how far 5 feet (or more) is in such situations. An AV should be able to calculate the parameters of and perform such a manoeuvre much more effectively than a human driver. However, what if the cyclist does something else, such as moving further into the carriageway during the overtake attempt? Should the AV try and pull back (even if it may not be safe to do so), give more way by entering further into the opposite lane (even though it may not be possible), and so on? This is but one example of a dilemma that can result from an AV performing a typically everyday action, with relatively low risk most of the time. That is, until something unexpected happens that in itself needs to be responded to and the outcome will not always be positive (e.g., entering further into to the opposite lane and then a car in that lane suddenly pulls out of a parking space, with it ultimately being too late for either party to avoid a collision).

Second, like any other aspects of legislation, to cope with complex and dynamically changing traffic environments, traffic laws are often deliberately written with a certain level of vagueness such that they can be flexibly applied to all situations (Endicott, 2001, 2011). Instead of providing specific requirements, clauses are usually principles that need to be adhered to as much as possible and are subject to interpretation and occasionally, discretion. For example, in the UK, the wording associated with what constitutes dangerous/careless driving is vague, which then demands evaluation on a case-by-case basis by drivers and law enforcers. For instance, driving a non-emergency services vehicle at 50 mph in a 30-mph speed limit zone is a violation of the ROTR, whereas driving at 33 mph in a 30-mph zone might be discretionary for a number of reasons that are not related to acting dangerously (e.g., such as in the overtaking-a-cyclist scenario described earlier – where the cyclist accelerates and it is no longer safe for the overtaking vehicle to move back into its lane).

Third, even in situations where the ROTR unequivocally prohibit certain behaviours, strictly sticking to traffic rules will not always or necessarily result in an optimal outcome. The emergency services vehicle scenario described earlier is an example. That is, if drivers do not take actions to provide a safe passing gap, the emergency services vehicle could be delayed, with a potentially life-threatening negative outcome. Another example is when there is an obstruction on the road ahead (e.g., a branch broken off a tree, a rubbish/trash bin/container blown over) due to weather or other circumstances. Vehicles should be able to circumvent the obstacle(s) or follow a diversion, even though such actions could be risky (e.g., crossing a broken white line in the middle of the road when it appears safe to do so) or perhaps even breaking a law if it prevents an incident (e.g., crossing solid double white lines in the middle of the road when it appears safe to do so). Inaction – i.e. stopping – would have a key consequence of potentially holding up a part of the traffic network in that area; potentially resulting in a standstill situation (with many examples of these for AVs to date).

Due to various limitations of the ROTR, existing traffic systems in all countries are effectively governed by a mixture of laws, advisory documents and cultural customs (Schmitt, 2020; Tennant et al., 2021). The inflexibility of a pure rule-based decision algorithm would risk removing the nuances and the adaptive aspects of driving which are instrumental in ensuring a smoother, efficient and on occasion, safer traffic environment. Therefore, desirable specifications for the driving behaviours of AVs need to possess a certain level of discretion in their judgements – just like human drivers have and exercise frequently. Deviations from the ROTR and sometimes traffic laws – via discretionary actions – are at times necessary and perhaps even advisable to ensure the optimal functioning of the traffic network (e.g., if a police officer signals to the driver of a vehicle to carefully make a turn even though the traffic light is on red).

2.4. Discretionary action

To accommodate the complexity of road traffic and to interact effectively with other road users (including cyclists, pedestrians, and so on), it is argued that a certain level of adaptability needs to be built into the behaviour specifications of AVs to optimise traffic flow, improve safety for all road users and pedestrians. We define discretionary actions as manoeuvres that a driver performs based on the evaluation of specific situations in order to maximise the safety of themselves and other road users and/or improve traffic flow. Legal discretionary actions include manoeuvres that are out of the ordinary but still within the boundary of laws. That is, they are permitted by law if certain conditions are satisfied. For example, in the UK, straddling broken white lines or white boxes with diagonal patterns are permitted only if it is to circumvent an obstruction, and the driver has deemed it safe to do so. Thus, legal discretionary actions represent a grey area where the vehicle operator needs to exercise judgement. They are only lawful when the decision(s) can be absolutely justified. In contrast, illegal discretionary actions involve manoeuvres that are explicitly prohibited by the national and/or local traffic laws. For example, in the UK, straddling solid double white lines or running a red light is illegal, regardless of the circumstances surrounding them. We argue that for AVs to be fully integrated into the current traffic systems and to be trusted by the public, discretionary actions that are calculated to be safe (based on the rich data available) should not only be allowed but also form part of their desirable behavioural specifications.

Programming AVs to enter into so-called grey areas or even to break the law may sound like a radical idea but the advocation (at least of the former) can be based on at least two grounds. First, human drivers already often do it. If AVs cannot match human drivers in their ability to be adaptive and exercise judgement, it will almost certainly result in traffic congestion (even gridlock situations), frustration of, and even danger to, other road users, and consequently loss of trust from users and observers. Second, with advanced sensors and decision algorithms that far exceed human cognitive capabilities (including perception, attention, judgement, and decision-making), it can be argued that it is far less risky for AVs to perform discretionary actions than human drivers. They will have not only calculated the safety parameters of the intended action based upon more information than a human driver would have and in faster time, but also considered potential unforeseen consequences of the action (e.g., that there is blind bend ahead and should something not go to plan during the action – a collision with a vehicle that may currently be out of sight could be possible).

However, a certain level of caution must be exercised before even attempting to program AVs to perform discretionary actions, or before developing standards and legal frameworks that might allow these to be implemented. Apart from tremendous technical difficulties associated with translating social codes and customs into programming code (including machine learning and AI) as well as setting decision-making thresholds, the psychological barriers are potentially colossal. Discretionary actions mean assuming an extra amount of risk – even for AVs that have cutting-edge technologies and are becoming even more advanced. Such risks need to be factored into standards, legislation, insurance cover, and so on. Actions always have consequences even when intentions are benign or benevolent. As argued earlier, discretionary actions might be less risky for AVs than for human drivers, but there will inevitably be

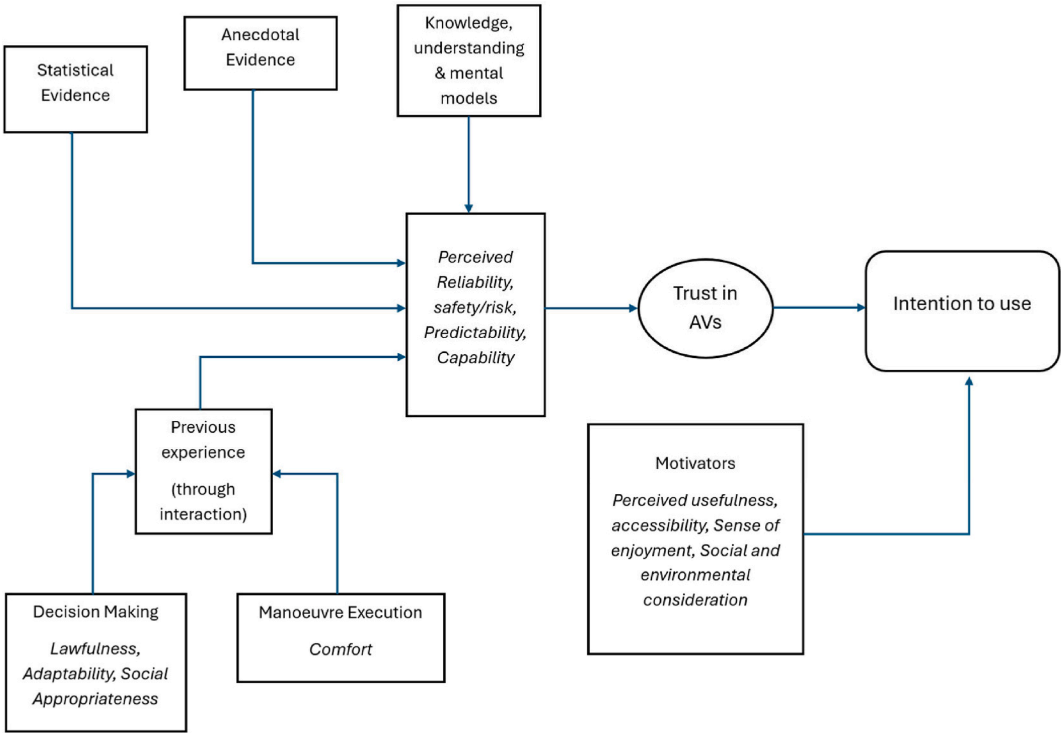

Statistical Evidence
Anecdotal Evidence
Knowledge, understanding & mental models
Perceived Reliability, safety/risk, Predictability, Capability
Trust in AVs
Intention to use
Previous experience (through interaction)
Motivators
Perceived usefulness, accessibility, Sense of enjoyment, Social and environmental consideration
Decision Making
Lawfulness, Adaptability, Social Appropriateness
Manoeuvre Execution
Comfort

Fig. 1. A conceptual framework for determinants of trust in AVs and intention to use.

occasions where the ultimate outcomes are negative. AVs can miscalculate and misread the behaviours of other road users, and discretion can thus potentially lead to collisions.

3. Contributions and research questions

Fig. 1 summarises the theoretical and empirical work reviewed in Section 2. Perceived reliability, safety, predictability, and capability are all important determinants of trust in AVs (Hancock et al., 2011, 2019; J. D. Lee & See, 2004; Naiseh et al., 2025; Parasuraman & Riley, 1997b), which can be informed through a wide range of channels including personal experience with the technology (Hoff & Bashir, 2015; Olaverri-Monreal, 2020), as a passenger in an AV, another road user interacting with the AV, or as a bystander who witnesses the AV negotiating its way through traffic. Apart from conveying a sense of reliability and safety, the behaviours of an AV also need to ensure comfort and safety in its execution of manoeuvres (Abe et al., 2015; Hartwich et al., 2018) as well as demonstrate lawfulness, social appropriateness, and adaptability in its decision-making (Bin-Nun et al., 2022; Tennant et al., 2021). Relating to this last point, while the capability of performing discretionary actions seems to be a feature that AVs need to possess to operate efficiently and further win user and/or public trust, it is imperative to understand the extent to which the public (including potential adopters) would accept these actions and the extra risk they entail. The current research serves as an initial step towards this understanding. A good indicator of public acceptance of AVs executing discretionary actions is the level of blame that they assign to an AV when such discretionary actions lead to a negative outcome, especially when compared to a human driver executing exactly the same actions. The judgement of blame has been investigated by a number of studies as an important indicator of attitudes towards AVs (e.g., Awad et al., 2020; Bennett et al., 2020; Hong, 2020; Hong et al., 2021; Liu & Du, 2021; Pöllänen et al., 2020; Wallbridge et al., 2024; Zhang et al., 2024). But research is lacking on blame attribution in the event of an accident following discretionary behaviours of AVs. The specific questions the current paper sets out to answer are:

a) Would AVs be blamed more than human drivers when they execute the same discretionary action in the same circumstances that lead to the same consequence?
b) To what extent is trust in an AV impacted in such events compared to a human driver?
c) How would witnessing an AV executing a discretionary action that leads to an accident or near-miss affect people's acceptance of self-driving technology in general?

4. Overview of the methodology

Three experiments investigated attitudes towards AVs executing (Experiments 1 and 2) or not executing (Experiment 3) discretionary actions compared to human drivers, after these actions (or inactions) were shown to have led to negative (i.e., traffic congestion or collisions) or positive outcomes (i.e., improved traffic flow or avoidance of collision). All three experiments employed a similar paradigm in which participants read multi-stage vignette stories featuring a "target vehicle", operated by either an advanced autonomous system or a highly capable human driver, performing (or opting to not perform – Experiment 3) a discretionary action. This entailed straddling the broken white line (Experiment 1) or double solid white lines (Experiments 2 and 3) in the middle of the road, that led to an acutely negative or more positive outcome across various scenarios. Participants' judgements of blame (with regard to the operator of the target vehicle and other parties involved) as well as their trust in the operator of the target vehicle (AV or human driver) were measured following each scenario. Also, possible changes in general acceptance of AVs were probed by measuring them both before and after having experienced all scenarios.

Negative outcomes preceded by discretionary actions are unique in the sense that they provide a salient candidate to feed mental mutation and counterfactual thinking (i.e., imagination of what might have been) (Roese & Olson, 1997), which might heighten the blame on the part of the driver. Research on counterfactual thinking has revealed that a negative event preceded by a novel or abnormal event will provoke more intense emotional reactions and harsher social judgements (Kahneman & Miller, 1986; Markman & McMullen, 2003; Miller et al., 1990) because it more readily induces what-if counterfactual thoughts (e.g., "the collision shouldn't have happened" or "the driver shouldn't have done that"). A discretionary action is by nature a deed that is out of the ordinary or out of convention which might intensify human rumination of how things could have been. Further, a discretionary action performed by an AV (e.g., straddling the double white line in the middle) might be seen as even more extraordinary or "abnormal" due to the fact that it violates the stereotypical perception of conventional computers and machines as rule-based systems for which improvisation should not be part of their behavioural characteristics (de Winter & Hancock, 2015; Fitts et al., 1951). Hence, it is predicted that AVs will be blamed more and trusted less than human drivers for executing the same discretionary actions under the same circumstances leading to the same outcomes because the former would provoke stronger counterfactual thoughts (e.g., "the vehicle should not have done it") which would lead to increased blame.

Experiment 3 complements the first two experiments by exploring the flip side of the same coin – people's attitudes towards AVs not executing discretionary actions (e.g., stopping – inaction, continuing to stay/drive in the same lane), even when they could and perhaps should have given the ultimate outcomes that vary in terms of acuteness (highly negative – e.g., colliding with a pedestrian; to less negative – e.g., causing congestion due to not overtaking a cyclist). It is expected that blame on "inaction" (as well as the judgement of trust) would display a reciprocal pattern to blame on "actions". That is, AVs will be blamed less and trusted more if they opted not to take a discretionary action compared to human drivers, even when such inactions lead to a negative outcome. Again, this is due to the perception that AVs, as machines, do not possess the capability to digress from the rules as human drivers do.

5. Experiment 1 & 2

These two vignette-based experiments were designed to compare attitudes – specifically in terms of trust and blame – towards AVs versus human drivers executing discretionary actions. The designs, procedures and stimuli (vignettes) of the two experiments were almost exactly identical, except for that the vignettes used in Experiment 1 all featured one *legal* manoeuvre − straddling *broken* white lines in the middle of the road − which is permitted by law but requires the operator of the vehicle to exercise discretion, whereas vignettes used in Experiment 2 featured the *illegal* variant of this manoeuvre − straddling *solid* white lines, which is explicitly prohibited by law. In both experiments, participants' judgements of trust and blame were probed after reading each of nine scenarios, which varied in other ways – e.g., some involved a response to an emergency situation (e.g., a pedestrian suddenly walking into the road ahead), others involved an environmental hazard (e.g., large body of water on the road ahead), and the remainder involved negotiating with other road users (e.g., a cyclist).

5.1. Method

5.1.1. Participants

One-hundred and twenty-nine participants were recruited to take part in Experiment 1 and 152 in Experiment 2 through the online participant recruitment platform *Prolific Academic* and were rewarded £3.75 each for a session lasting approximately 30 min. As the pre-screening criteria dictated, participants were all over 18 years of age with self-reported normal or corrected-to-normal vision and all were UK residents, either with English as a first language or highly proficient in English as a second language. In Experiment 1, the sample consisted of 81 females (63%) and 48 males (37%), with a mean age of 36.43 ($SD = 15.76$, $Min = 18$, $Max = 75$). One-hundred participants (77.5% of the sample) had a full driving license at the time of the experiment. Of those, the average number of years of driving experience was 20.92 ($SD = 15.96$, $Min = 1$, $Max = 61$) and the average annual mileage of driving was 5600.42 ($SD = 4716.03$, $Min = 0$, $Max = 25{,}000$). In Experiment 2, the sample consisted of 69 females (45.4%) and 83 males (54.6%), with a mean age of 36.43 ($SD = 15.76$, $Min = 18$, $Max = 75$). One-hundred and five participants (69%) had a full driving license. Of those, the average number of years of driving experience was 24.83 ($SD = 14.80$, $Min = 1$, $Max = 59$) and the average annual mileage of driving was 6626.95 ($SD = 4868.76$, $Min = 0$, $Max = 25{,}000$).

5.1.2. Materials

In Experiment 1, experimental stimuli and questions were delivered via an online questionnaire created and hosted in *Qualtrics*. The main stimuli were nine vignette stories/scenarios in which participants were asked to focus on the behaviour of nine different "target" vehicles (labelled in the scenarios as Vehicle A, B, C…to I with randomized order). All scenarios featured a target vehicle executing the same generally legal discretionary action – straddling a broken white line in the middle of the road. However, these actions were induced under different driving situations. Some were induced by an emergency as the result of another road user violating the traffic rules (for example, in the 'Pedestrian' scenario: a pedestrian was said to suddenly step into the road in front of the target vehicle) whereas the others were induced by less acute situations (for example, in the 'Ambulance' scenario: the target vehicle was said to be stuck in a traffic congestion when the crew of the ambulance behind sounded the siren, flashed the blue lights, and tried to get through the busy traffic). See Table 1.1 for a summary of the premises of all scenarios. Scenarios were designed to capture a wide range of traffic situations – varying in acuteness, under which a discretionary action (i.e., choice to cross the broken white lines or to stop) might be necessary and possible.

All vignettes in Experiment 1 consisted of two elements: a one-page introduction and a story section. The introduction guided participants on how to identify the target vehicle they were to focus on (i.e., "a white vehicle with an orange "eye" symbol on its roof"). They were instructed that the scenario had three parts and they had 30 s to view each part, after which the page would automatically advance to the next part. The time constraint was introduced in order to minimise the variation among participants in the level of processing experimental materials (e.g., to prevent some participants from fixating on one detail of the scenario and spending considerably more time on this aspect of the story than the others). Thirty seconds were shown by our pilot studies to be a reasonable amount of time to read and understand each part of the story.[1] Participants were also instructed that each scenario part would only be shown to them once and they would be unable to return to scenario parts after or to re-review any aspects of the scenario after it had ended. This constraint was again introduced to control for the level of information processing and time spent viewing the materials. The story section was broken into three parts, presented sequentially on separate slides (see Fig. 2.1 − full scenarios can be found in Appendix B). Each part of the main story consisted of a textual narrative occupying the top half of the page and a corresponding pictorial illustration occupying the bottom half. Labels and legends were provided to assist comprehension.

Part 1 described the "premise" of the story − the traffic situation that the target vehicle was facing. This was accompanied by a passage from the UK Highway Code (the ROTR in the UK) which was included to clarify, inform and (for drivers and learner drivers) remind participants what is legal and what is not. Part 2 described the action that the target vehicle (AV or the driver of a non-AV) chose to take, which in all scenarios was to cross the broken white line in the middle of the road to e.g., avoid colliding with the

[1] The results of a pilot study (N = 6) without the time constraint showed that the average time spent on each part of the vignette was 24.99 s after removing outliers (Min = 5.22, Max = 95.48, SD = 17.93). A second pilot study (N = 40) compared the ratings of main dependent variables between a group (n = 20) with the 30-second restriction with another group (n = 20) without. The ratings of the two groups did not significantly differ from each other, indicating the time constraint does not affect participants' understanding of the vignettes and their responses to the questions.

Table 1.1
Summary of premises of the nine scenarios.

Scenario Label	Premise
Pedestrian	Vehicle A is driving at the speed limit on a two-lane road (one lane in each direction). A pedestrian suddenly walks into the road in front of Vehicle A.
Ambulance	Vehicle B is stuck in a traffic jam on a four-lane road (two lanes in each direction). Vehicle B is queuing in the central-left lane. Suddenly the driver of the ambulance on its left puts its siren and flashing lights on.
Bus	Vehicle C is driving at the speed limit on a two-lane road (one lane in each direction). The bus in front of Vehicle C pulls left into a designated bus stop bay, clearing the way for Vehicle C. Vehicle C goes ahead to continue its journey. Before Vehicle C clears the length of the bus, a pedestrian suddenly walks out from in-front of the bus to cross the road.
Bin	Vehicle D is driving at the speed limit on a two-lane road (one lane in each direction). It is a windy day. A rubbish bin is blown from the driveway into the road and falls over, blocking the path of Vehicle D.
Tow Truck	Vehicle E is driving at the speed limit on a two-lane road (one lane in each direction). There are parked vehicles on the right-hand side of the road. The tow truck in front of Vehicle E stops and turns on its hazard lights, blocking Vehicle E and the traffic behind.
Trailer	Vehicle F is driving at the speed limit on a two-lane road (one lane in each direction). It is following a vehicle towing a trailer. The trailer becomes detached from the tow bar of the vehicle due to a failure of the fixing. The trailer does not have a safety chain as a secondary fixing.
Puddle	Vehicle G is driving at the speed limit on a two-lane road (one lane for either direction) through a rocky area. It gets over the top of a hill. A large deep puddle blocks its path.
Deer	Vehicle H is driving at the speed limit on a two-lane road (one lane in each direction). A deer suddenly walks out in front of the vehicle from the left side of the road.
Cyclist	Vehicle I is driving at the speed limit on a two-lane road (one lane in each direction). There is a cyclist riding in front close to the kerb, slowing down Vehicle I and the traffic behind.

pedestrian, overtake a cyclist, and so on. Again, this was accompanied by a legal passage from the Highway Code. Part 3 described the outcome of the scenario. Depending on conditions, there were two alternative versions of outcomes – versions which featured a subsequent accident (e.g., Vehicle X collides with a car backing out of a driveway in the opposite lane) and versions which featured near-misses (e.g., Vehicle X nearly collides with the reversing car). Regardless of version type, part 3 always consisted of two subsections: The first paragraph described the *factual* – what happened following the target vehicle performing the discretionary action, whereas the second paragraph described the *counterfactual* – what would have happened if the target vehicle had not performed the discretionary action. A factual outcome of an accident (i.e., a collision) was always coupled with an *upward* counterfactual outcome about how things could have been *better* (Markman et al., 1993; Roese, 1994). Conversely, a near-miss outcome was always coupled with a *downward* counterfactual outcome about how things could have been *worse*.

The “counterfactual” components were crucial to the manipulation of outcome valence because the subjective experience of the valence of an event is highly susceptible to comparison anchors (Helson, 1964; Lazarus, 1991). In the context of traffic incidents, the appraisal of the desirability of an event will be greatly affected by the imagination of alternative reality of what might or could have happened. Hence it was deemed natural and sensible to include counterfactual information as part of the variable manipulation.

Materials were exactly the same in Experiment 2 except for that in all vignette scenarios, the broken white lines in the middle of the road were replaced by solid double white lines, both in the text and the pictorial illustrations. A “No Overtaking” sign was added before the area where the target vehicle performed the discretionary action (see Fig. 2.2). Full scenarios can be found in Appendix C.

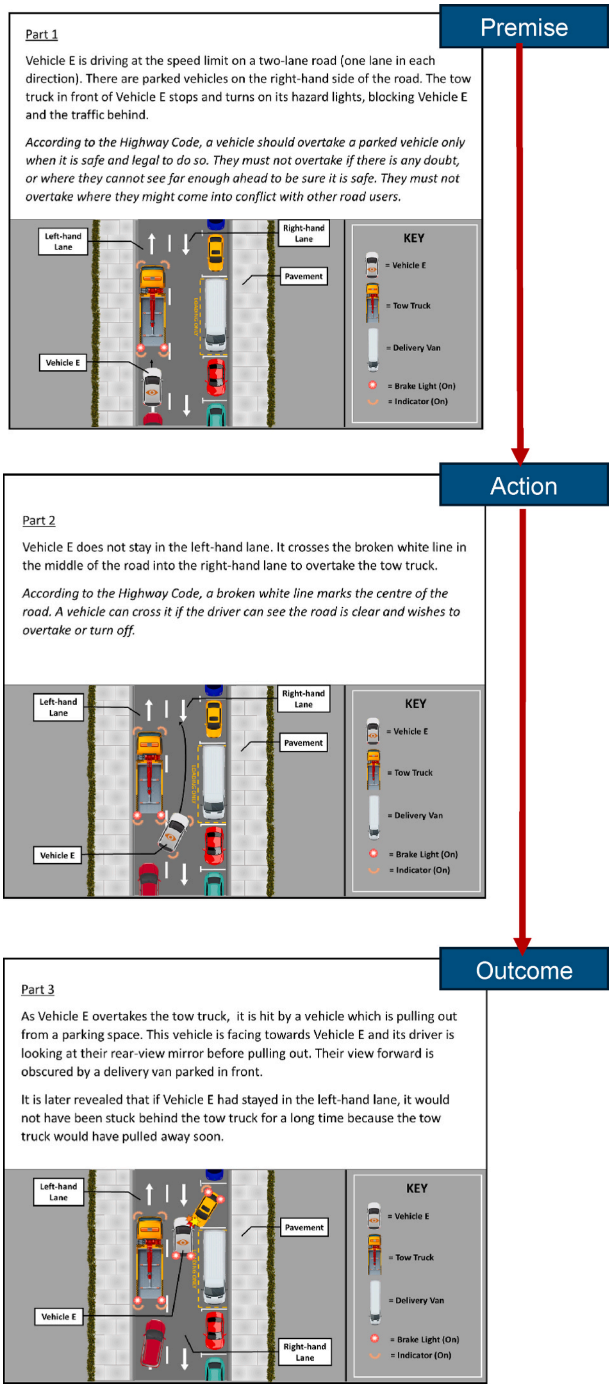
5.1.3. Design

Both experiments adopted a 2 (Operator: autonomous system (AS), human driver) X 2 (Outcome: accident, near-miss) between-participant design. This design was chosen to minimise possible carryover effects and the probability that participants could guess the research hypotheses due to exposure to multiple treatments. Table 1.2 summarises the number of participants in every condition. Each participant was presented with the nine vignettes in a random order featuring different driving situations. They were asked to answer questions that followed every vignette. Operator was manipulated by informing the participants the target vehicle in every scenario was either operated by a human driver (HD condition) or an autonomous system (AS condition). Outcome was also manipulated between-participant by presenting alternative endings of the vignette stories. In the accident outcome condition, all vignettes concluded with a collision resulting in minor injuries. In contrast, in the near-miss outcome condition, stories concluded with a near-collision with no tangible consequences.

Three main dependent variables were measured after each scenario across both experiments: (i) blame on the operator of the target vehicle; (ii) blame on the third parties involved (Third Party A – the party whose action directly induces the manoeuvre being executed – e.g., the pedestrian stepping into the road; and Third Party B – the party whose action directly contributes to the final outcome – e.g., the driver of the vehicle reversing out of the driveway and colliding, or nearly colliding, with the target vehicle); and (iii) the degree of trust in the operator of the target vehicle. In addition to these incident-specific variables, participants’ general acceptance of AVs in the AS condition (including general trust in and the likelihood of using AVs in the future) was measured at the outset of the experiment (before any scenarios were experienced) as well as in the end (after all scenarios had been presented and scenario-specific dependent variables recorded).

5.1.4. Procedure

The two experiments shared the exact same procedure. Upon accepting the invitation from *Prolific*, participants were provided with a link through which they accessed the online *Qualtrics* questionnaire. The first page of the questionnaire was an information sheet



Part 1

Vehicle E is driving at the speed limit on a two-lane road (one lane in each direction). There are parked vehicles on the right-hand side of the road. The tow truck in front of Vehicle E stops and turns on its hazard lights, blocking Vehicle E and the traffic behind.

According to the Highway Code, a vehicle should overtake a parked vehicle only when it is safe and legal to do so. They must not overtake if there is any doubt, or where they cannot see far enough ahead to be sure it is safe. They must not overtake where they might come into conflict with other road users.

Part 2

Vehicle E does not stay in the left-hand lane. It crosses the broken white line in the middle of the road into the right-hand lane to overtake the tow truck.

According to the Highway Code, a broken white line marks the centre of the road. A vehicle can cross it if the driver can see the road is clear and wishes to overtake or turn off.

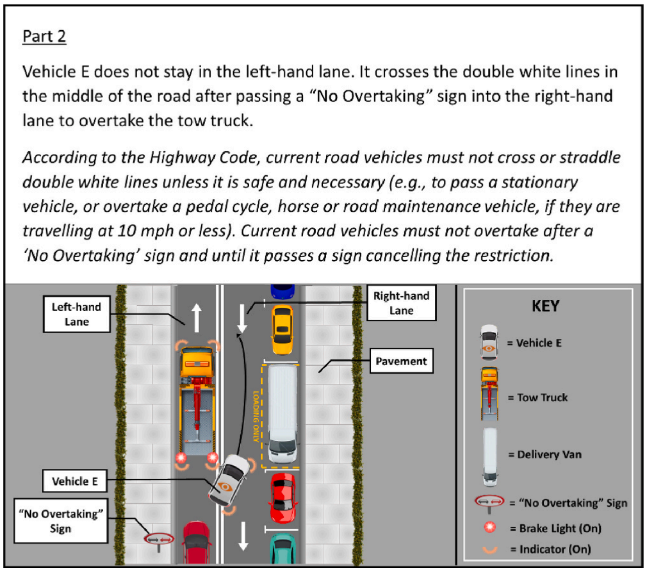
Part 3

As Vehicle E overtakes the tow truck, it is hit by a vehicle which is pulling out from a parking space. This vehicle is facing towards Vehicle E and its driver is looking at their rear-view mirror before pulling out. Their view forward is obscured by a delivery van parked in front.

It is later revealed that if Vehicle E had stayed in the left-hand lane, it would not have been stuck behind the tow truck for a long time because the tow truck would have pulled away soon.

Fig. 2.1. Structure and flow of the nine vignettes in Experiment 1. *Note.* The example features a scenario in which the target vehicle needs to negotiate a parked tow truck stopped in the road ahead. In this example, the scenario ends with an accident (i.e., collision).

which stated that the online experiment formed part of a research project investigating people's opinions of scenarios involving road vehicles, other road users and traffic infrastructure and it would involve them reading a series of short stories featuring various traffic situations and responding to questions relating to their opinions of the story and driving in general. This was followed by informed consent. After this, participants were directed to questions concerning age, gender and driving experience (with "prefer not to answer" options). Participants in the AS condition were then presented with two questions regarding their general acceptance of AVs where they responded by rating on two 11-point scales, one concerning their likelihood of using an AV when the technology becomes available (*Imagine that fully autonomous vehicles will be deployed on a large scale on UK roads within the next 12-months. Please rate how*



Part 2

Vehicle E does not stay in the left-hand lane. It crosses the double white lines in the middle of the road after passing a "No Overtaking" sign into the right-hand lane to overtake the tow truck.

According to the Highway Code, current road vehicles must not cross or straddle double white lines unless it is safe and necessary (e.g., to pass a stationary vehicle, or overtake a pedal cycle, horse or road maintenance vehicle, if they are travelling at 10 mph or less). Current road vehicles must not overtake after a 'No Overtaking' sign and until it passes a sign cancelling the restriction.

Fig. 2.2. Example of scenario featuring an illegal discretionary action. *Note.* Compared to Experiment 1, the broken white line in the middle of the road has been replaced by solid double white lines. There was also a "No Overtaking" road sign added to the pavement on the left side of the street.

Table 1.2
Condition allocations of participants in Experiment 1, 2 & 3.

Condition	AS	HD	Total
Collision	32/36/43	33/38/42	65/74/85
Near-miss	33/40/43	31/38/43	64/78/86
Total	65/76/86	64/76/85	129/152/171

Note. The numbers in each cell correspond to Experiment 1, 2 and 3 respectively.

likely you would be to use an autonomous vehicle on a scale from 0 – Extremely unlikely to 10 – Extremely likely), and the other concerning their trust in the AV technology (*Imagine that fully autonomous vehicles will be deployed on a large scale on UK roads within the next 12-months. Please rate how much you trust autonomous vehicle technology on a scale from 0 – do not trust at all to 10 – completely trust*).

Participants in both conditions were presented with general instructions regarding reading the scenarios and responding to the questions (Samples of instructions can be found in Appendix A). Crucially, instructions for the AS condition stated that in every scenario, the target vehicle was driven by "*a Level-5 autonomous system (*SAE, 2021*) with an impeccable safety record*", whereas the instructions for the HD condition stated that the vehicle was driven by *"a professional driver with an impeccable safety record"*.

Participants in both experiments then completed one practice trial before working through – one-by-one the nine vignette scenarios. After each scenario, judgements of blame on the operator of the target vehicle were measured by an 11-point scale (e.g., "*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?*") with 0 representing "not at all" and 10 representing "completely". Judgements of blame on the third parties that were involved in the incident were measured in a similar way (e.g., "*Based on the scenario you just experienced, to what extent do you think the pedestrian should be blamed for the incident that just took place?*"). Trust in the operator/driver of the target vehicle was measured by an 11-point scale (e.g., "*Based on the scenario you just experienced, how much would you trust the driver of/the autonomous system that controls Vehicle X to operate safely on the road in the future?*") with 0 representing "would not trust them at all" and 10 representing "would completely trust them". The order of the questions was randomized to minimise possible systematic carry-over effects from answering questions in the same order. Participants also answered attention-check questions: involving selection of a particular number on an 11-point scale (e.g., "*If you are reading this, please select 6.*").

After completing the scenarios, participants' general acceptance of AVs was measured again in both experiments using the same set of questions described above. They were then debriefed and given an opportunity to leave comments via a text box.

5.2. Results and discussion

The correct rates in responses to the attention questions in both experiments were 100%, indicating that all participants acted diligently and attentively when undertaking the experiment. In order to gain insights into the similarities and differences among the nine scenarios used, a factor analysis was conducted with a Principle Components Analysis (PCA) and a rotation method of Varimax with Kaiser normalization on participants' ratings of blame on the target vehicle in Experiment 1, which revealed a two-factor solution:

scenarios Pedestrian, Bus, Trailer, Puddle and Deer latched strongly to Factor 1 whereas scenarios Ambulance, Bin, Tow Truck and Cyclist latched on Factor 2. Factor 1 scenarios seem to feature more acute situations or emergencies where decisions are forced and need to be made quickly given that the stake of inaction is high (e.g. an almost certain collision and accident). In contrast, Factor 2 scenarios seem to feature situations with low acuteness where decisions are voluntary (e.g., to overtake a cyclist) and are not overly time sensitive (apart from e.g., taking longer to make a journey, traffic congestion building, and so on). Hence, we grouped scenarios in both Experiments into two categories accordingly and created a new repeated measures independent variable – Situation Acuteness with two levels: low and high.

5.2.1. Blame

Figs 2.3a and 2.3b illustrates mean ratings of blame on Vehicle X and Parties A and B in Experiment 1 (featuring legal discretionary actions). Levels of blame on the target vehicle were relatively low. The vast majority of the mean ratings fell below the mid-point of the scale within both the accident and near-miss outcome situations. Also, blame on the target vehicle appears lower than blame on Third Party B (whose action immediately precedes the collision or the near-miss incident) in all cases and lower than Third Party A (whose action induces the discretionary action of the target vehicle) when situation acuteness was high. With regard to the comparison between the two third parties, and for the high acuteness scenarios, Third Party A seemed to receive more blame in general than Third Party B. By comparison, blame on the target vehicle in Experiment 2 was in general higher than in Experiment 1 (see Figs. 2.4a and 2.4b), which is understandable since Experiment 2 featured illegal manoeuvres. This increase was especially visible in the Outcome: accident condition, where blame on the target vehicle operated by an autonomous system (AS) was similar to or higher than both third parties on the whole. The following analyses will focus on blame on the target vehicle, which is of the primary interest of the current paper.

The two experiments displayed similar patterns as to how blame ratings were affected by Outcome Severity. A 2 (Operator) X 2 (Outcome) X 2 (Situation Acuteness) mixed ANOVA[2] reveals a significant main effect of Outcome in both experiments ($F(1, 125) = 33.29$, $p < 0.001$, $\eta^2 = 0.21$; $F(1, 148) = 74.95$, $p < 0.001$, $\eta^2 = 0.34$, respectively) on the level of blame on the target vehicle: mean blame was higher when the outcome was a collision compared to a near-miss ($M = 3.92$, $SE = 0.22$ versus $M = 2.15$, $SE = 0.22$ in Experiment 1; $M = 5.49$, $SE = 0.22$ versus $M = 2.90$, $SE = 0.21$ in Experiment 2). However, and as Figs. 2.5 and 2.6 illustrates, the magnitude of this difference seems to be dependent on the acuteness of the situation, supported by a significant Outcome X Situation Acuteness interaction in both experiments ($F(1, 125) = 36.46$, $p < 0.001$, $\eta^2 = 0.23$; $F(1, 148) = 19.90$, $p < 0.001$, $\eta^2 = 0.12$, respectively). That is, in Experiment 1, as the situations became less acute, the difference in blame between the two outcome conditions reduced from a significant ($p < 0.001$) to a non-significant level ($p = 0.024$). A similar pattern was observed in Experiment 2 – although the differences in blame between the two outcome conditions were both still significant ($ps < 0.001$). These patterns seem to be the result of the opposing effects of situation acuteness on blame in the two outcome conditions. That is, when the outcome was a near-miss, blame in the low acuteness conditions was higher than that in the high acuteness conditions ($p = 0.002$ in Experiment 1 and $p < 0.001$ in Experiment 2). However, this pattern did not occur when the outcome was a collision (reversed in Experiment 1 ($p < 0.001$), non-significant in Experiment 2 ($p = 0.255$)).

The two experiments exhibited different patterns as to the effect of Operator Type on blame and how this was moderated by Situation Awareness. In Experiment 1, the main effect of Operator was marginally non-significant ($F(1, 125) = 3.86$, $p = 0.052$). However, there was a significant Operator X Situation Acuteness interaction ($F(1, 125) = 5.72$, $p = 0.018$, $\eta^2 = 0.04$). Bonferonni post-hoc comparisons reveal that when situation acuteness was high, autonomous systems were blamed more than human drivers ($p = 0.005$). However, there was a non-significant difference when situation acuteness was low ($p = 0.535$). The main effect of Situation Acuteness and all interactions among the three independent variables other than the ones listed above were non-significant ($p > 0.05$).

Hence, and in contrast to our hypothesized monotonic pattern that blame should be higher for AVs in all scenarios, the results in Experiment 1 indicate a more nuanced picture in which judgement of blame was moderated by the contextual factor of situation acuteness. A possible explanation is that participants' focus of judgement has varied depending on the level of acuteness of the situation. It should be noted that our hypotheses were based on the assumption that judgement of blame would be based on consideration of *whether* the discretionary action should have been executed. However, the consideration of *how well* the discretionary action was executed could also have informed blame. When acuteness was high, performing the discretionary action would have appeared to be more justifiable, therefore the focus of judgement would be more likely to be on *how well* the action had been performed, whereas when the acuteness is low, the focus of judgement would be more likely to be on the decision of discretionary action itself – i.e., whether it had been sensible to perform the action in the first place. It was not specified to participants which aspects of the driver behaviours they should base their blame judgements. Therefore the "*how well*" consideration, which is more associated with the expectation of the driver's competence and capability, might well have affected their blame ratings.

This speculation is supported by the pattern of results being consistent with the proposition that the perception of the capabilities of AVs compared to human drivers is context-specific (see Zhang et al., 2024). That is, AVs are perceived to be better than human drivers

[2] Before conducting the main analyses, the potential effect of individual differences in driving experience on blame ratings was explored in both experiments by A 2 (Driving Experience: qualified driver, non-qualified driver) X 2 (Operator: autonomous system, human driver) X 2 (Outcome: accident, near-miss) X 2 (Situation Acuteness: high, low) mixed ANOVAs, which revealed that being a qualified driver (having driving experience) had no significant effect on blame rating in either experiment ($F(1, 121) < 0.01$, $p = 0.982$; $F(1, 144) = 1.98$, $p = 0.162$, respectively). There was also no significant interaction between driving experience and any other main independent variable ($ps > 0.05$). Hence, driving experience was excluded from further analyses.

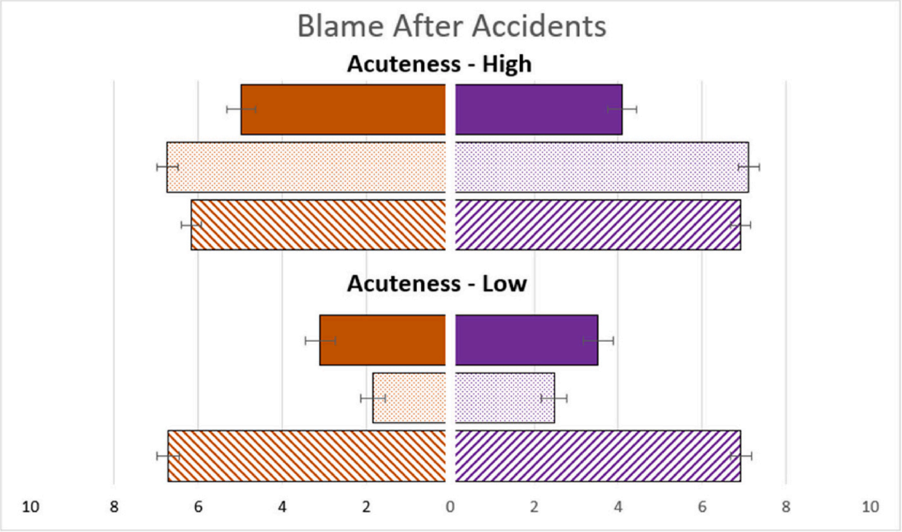


Fig. 2.3a. Mean ratings of blame on the target vehicle and third parties in the Outcome: collision condition in Experiment 1 (Error bars = ±1 SE).

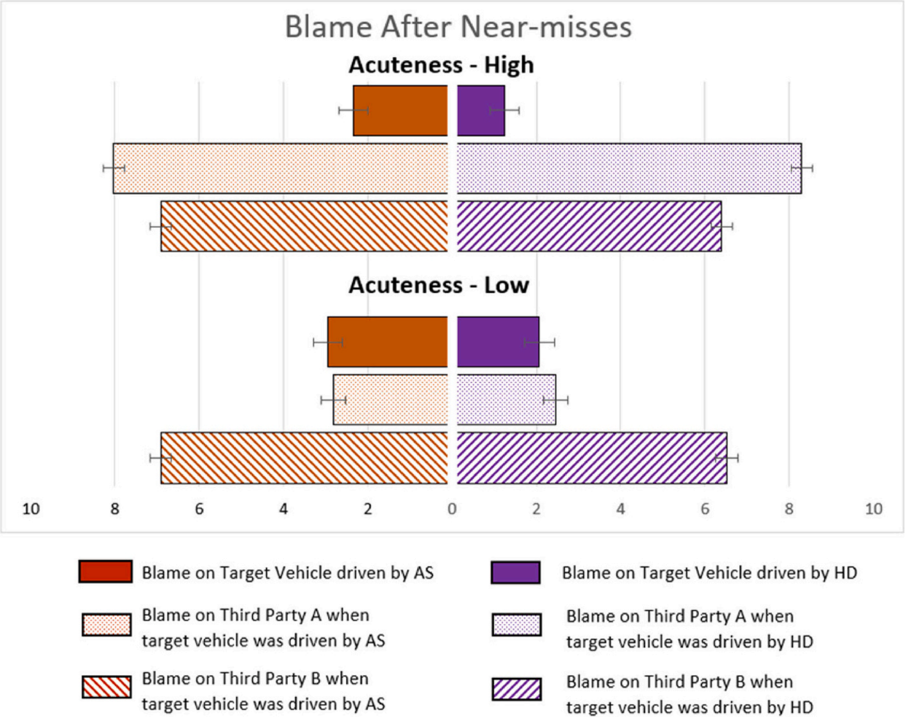


Fig. 2.3b. Mean ratings of blame on the target vehicle and third parties in the Outcome: near-miss condition in Experiment 1 (Error bars = ±1 SE).

at driving tasks that require speed and precision, but worse than human drivers at tasks that require inductive reasoning and judgement. Hence, it seems AVs were blamed more than human drivers in high acuteness situations because they were expected to react faster and possess more computing power to plot a safe solution. However, they were blamed less than humans in low acuteness situations because they were expected to be less capable of "reading" a situation and exercise improvision. It is speculated that this process has at least partially influenced the judgement process of our participants in Experiment 1 to result in the current pattern. This speculation also received some support from the significant Outcome X Situation Acuteness interaction described earlier – that fact that blame ratings were more affected by the outcome valence when the situation acuteness was high compared to low.

In contrast, the pattern of results in Experiment 2 demonstrated a more monotonic effect of Operator type on blame, regardless of situation acuteness. In Experiment 2, The main effect of Operator was significant ($F(1, 148) = 5.12$, $p = 0.025$, $\eta^2 = 0.03$). In general,

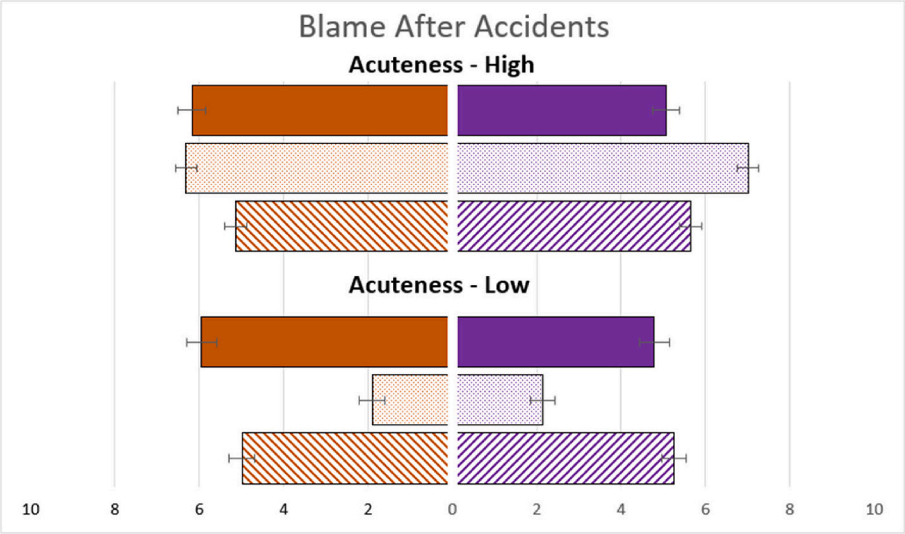


Fig. 2.4a. Mean ratings of blame on the target vehicle and third parties in the Outcome: accident condition in Experiment 2 (Error bars = ±1 SE).

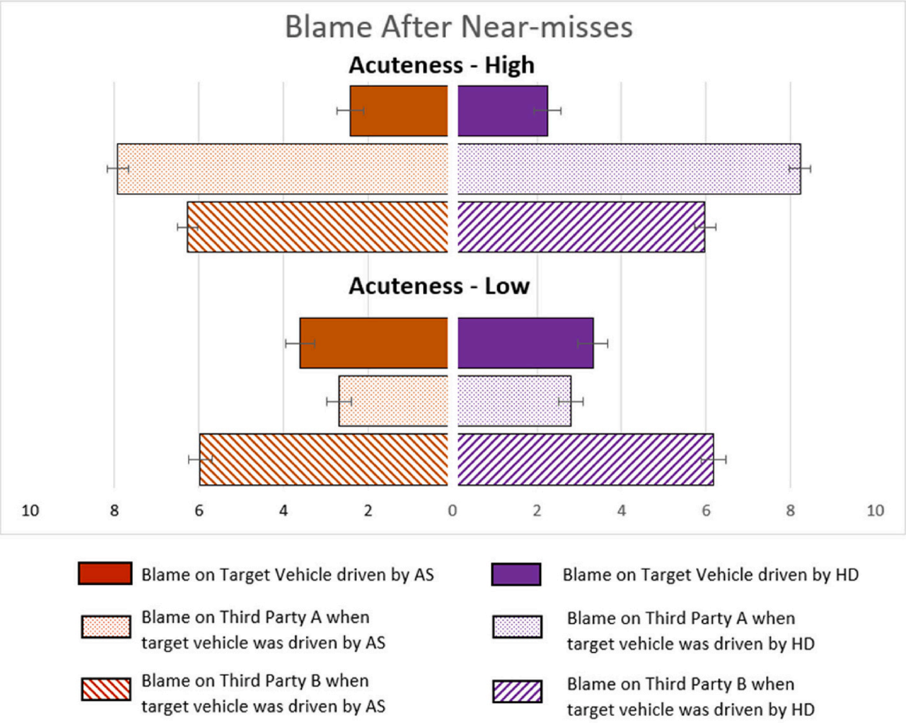


Fig. 2.4b. Mean ratings of blame on the target vehicle and third parties in the Outcome: near-miss condition in Experiment 2 (Error bars = ±1 SE).

blame was higher when the vehicle was operated by an autonomous system ($M = 4.54$, $SE = 0.21$) than a human driver ($M = 3.86$, $SE = 0.21$) (See Fig. 2.2). This effect was consistent across scenarios of different situation acuteness levels evidenced by a lack a significant interaction between Operator and Situation Acuteness $F(1, 148) = 0.08$, $p = 0.777$). All interactions among the independent variables were non-significant ($p > 0.05$). This pattern is consistent with our initial hypothesis of a uniform effect of operator on blame but differs from Experiment 1, in which blame on autonomous systems was only higher than human drivers in Situation Acuteness High condition. This change of pattern might reflect a shift in the judgement process when actions become illegal. That is, the violative nature of the illegal actions featured in Experiment 2 might have shifted the focus of judgement from *how well* the action was executed to *whether* it should have been executed at all in the first place. Perhaps then, fundamental stereotypical perceptions of machines (and

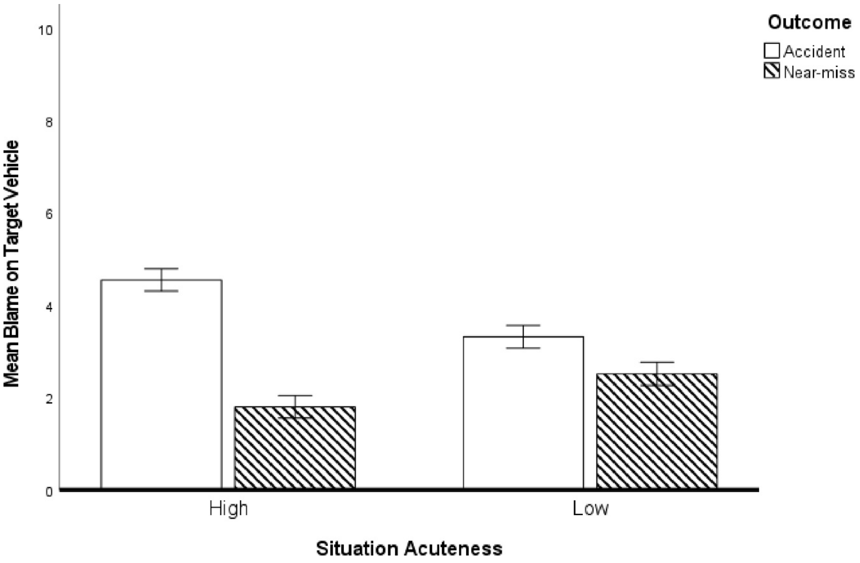


Fig. 2.5. Mean ratings of blame on the target vehicle in scenarios of different situation acuteness across two outcome conditions in Experiment 1 (Error bars = ±1 SE).

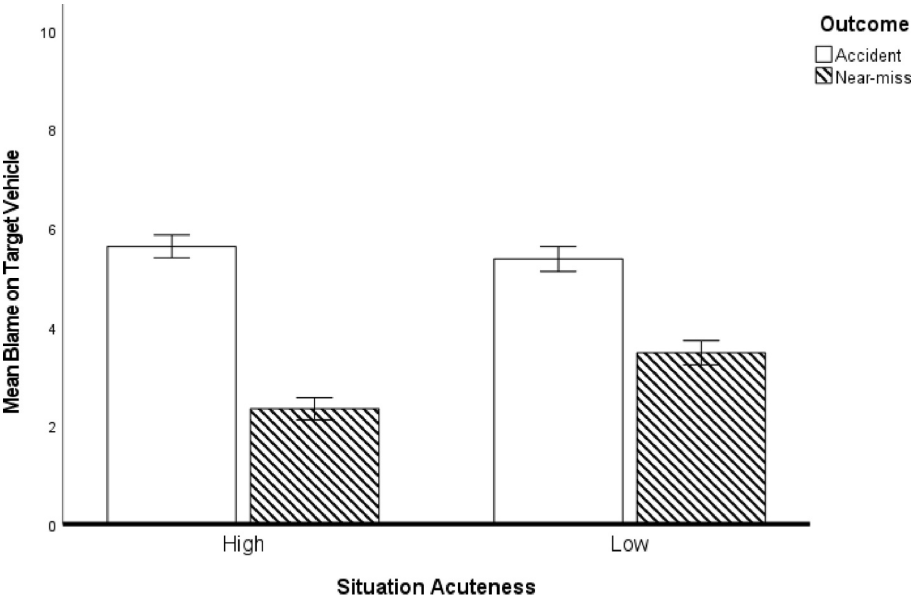


Fig. 2.6. Mean ratings of blame on the target vehicle in scenarios of different situation acuteness across two outcome conditions in Experiment 2 (Error bars = ±1 SE).

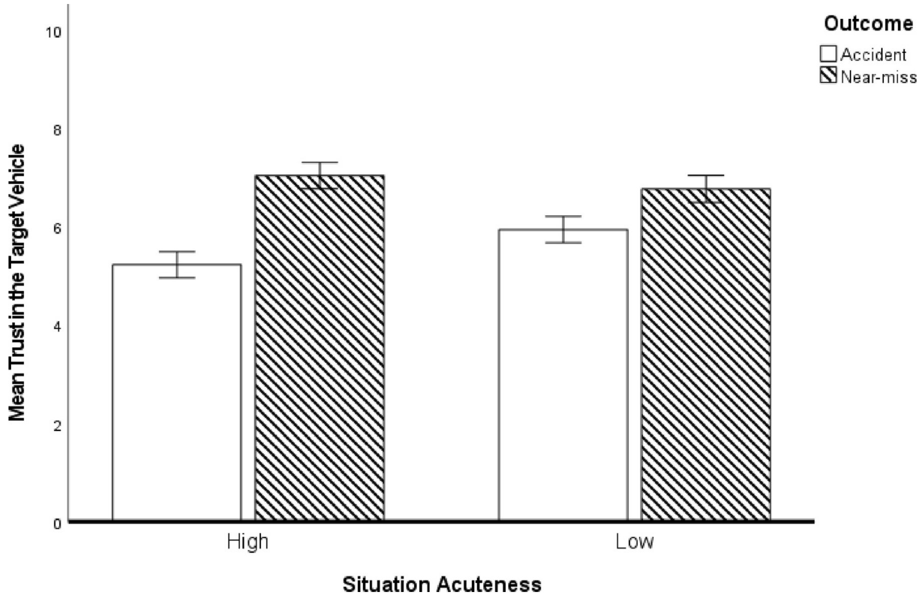


Fig. 2.7. Mean ratings of trust in the target vehicle in scenarios of different situation acuteness across two outcome conditions in Experiment 1 (Error bars = ±1 SE).

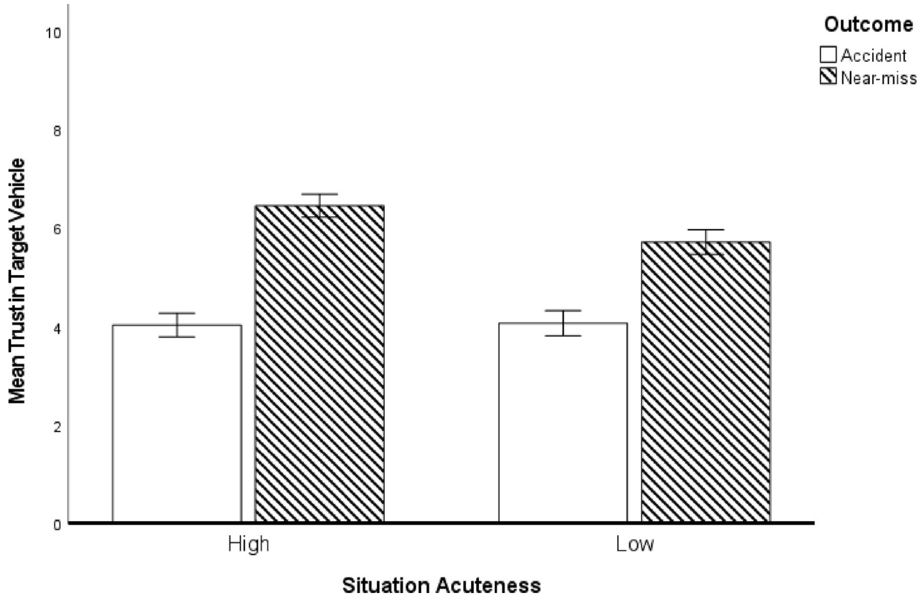


Fig. 2.8. Mean ratings of trust in the target vehicle in scenarios of different situation acuteness across two outcome conditions in Experiment 2 (Error bars = ±1 SE).

AI) as being rule-based feeds into expectations that they should not violate laws: they receive a high level of blame if they do.

5.2.2. Operator-specific trust

Patterns were similar between the experiments, although the overall level of trust was lower in Experiment 2 than in Experiment 1 (see Figs. 2.7 & 2.8). From a 2 (Operator) x 2 (Outcome) x 2 (Situation Acuteness) ANOVA,[3] a significant main effect of Situation Acuteness was found in both experiments ($F(1, 125) = 4.89, p = 0.029, \eta^2 = 0.04$; $F(1, 148) = 9.33, p = 0.003, \eta^2 = 0.06$. respectively) as well as Outcome ($F(1, 125) = 12.97, p < 0.001, \eta^2 = 0.09$; $F(1, 148) = 38.53, p < 0.001, \eta^2 = 0.21$, respectively) on trust in the driver/operator of the target vehicle. The two variables also significantly interacted ($F(1, 125) = 23.93, p < 0.001, \eta^2 = 0.16$; $F(1, 148) = 11.43, p < 0.001, \eta^2 = 0.07$). As Figs. 2.7 and 2.8 illustrates, when the situation acuteness was high, a near-miss outcome induced a higher level of trust than the accident outcome in both experiments ($ps < 0.001$). This difference reduced (but was still significant) when situation acuteness was low ($p = 0.033; p < 0.001$). There was also a significant main effect of Operator in both Experiments ($F(1, 125) = 33.46, p < 0.001, \eta^2 = 0.21$; $F(1, 148) = 46.62, p < 0.001, \eta^2 = 0.24$, respectively). Participants trusted human drivers more than autonomous systems ($M = 7.29, SE = 0.26$ versus $M = 5.17, SE = 0.26$ in Experiment 1; $M = 6.17, SE = 0.23$ versus $M = 3.94, SE = 0.23$ in Experiment 2). This is consistent between high and low acuteness scenarios, evidenced by a lack of significant interaction between Situation Acuteness and Operator in both experiments ($F(1, 125) = 0.25, p = 0.616$; $F(1, 148) = 0.05, p = 0.821$). All other main effects and interactions were non-significant ($p > 0.05$).

Overall, ratings of operator-specific trust exhibited a general favor towards human drivers, which was not a perfect reciprocal pattern of the blame findings, especially those in Experiment 1. Correlational analyses revealed a strong negative relationship between operator-specific trust and blame overall in both experiments but the strength of correlation was larger in the HD than in the AS condition ($r = -0.77, p < 0.001$ versus $r = -0.58, p < 0.001$ in Experiment 1; $r = -0.81, p < 0.001$ versus $r = -0.61, p < 0.001$ in Experiment 2). In the AS condition in particular, operator-specific trust was strongly correlated with pre-existing acceptance (i.e., general trust in AVs and likelihood of adopting AVs) (both $rs = 0.61, p < 0.001$ in Experiment 1 and $r = 0.44$ and $0.57, ps < 0.001$, in Experiment 2) whereas there was no significant correlation between blame and pre-existing acceptance ($r = -0.13, p = 0.317$; $r = -0.11, p = 0.370$, respectively in Experiment 1and $r = -0.10, p = 0.390$; $r = -0.15, p = 0.199$, respectively in Experiment 2). These findings indicate that although judgement of blame might inform the operator-specific trust, it is not the only factor – operator-specific trust is also hugely influenced by pre-existing acceptance of AVs in general. This could have given operator-specific trust a somewhat more "stable" attribute, compared to the judgement of blame, which could be more susceptible to incident-specific factors.

5.2.3. General trust towards AVs and likelihood of adoption

In the AS condition of both experiments, participants' general acceptance of AVs were measured both before and after reading the nine scenarios. Ratings of trust in AVs in general were analyzed using a 2 (Stage: pre-trial, post-trial) X 2 (Outcome: accident, near-miss) mixed ANOVA, which revealed a significant main effect of Stage in Experiment 1 ($F(1, 63) = 12.63, p < 0.001, \eta^2 = 0.17$). As Fig. 2.9a illustrates, participants' trust in AVs improved after witnessing multiple trials of AVs executing legal actions ($M = 3.80, SE = 0.34$ pre-trial versus $M = 4.69, SE = 0.36$ post-trial). However, the magnitude of improvement seems to depend on the outcome of the incident ($F(1, 63) = 4.72, p = 0.034, \eta^2 = 0.07$) and the difference was only significant when the legal discretionary action led to a near-miss ($p < 0.001$) instead of an accident ($p = 0.336$). In Experiment 2, the main effect of Stage was not significant ($F(1, 74) = 0.002, p = 0.966$). However, as in Experiment 1, the effect of witnessing AVs performing illegal discretionary actions on general trust was found to be dependent on Outcome ($F(1, 74) = 5.15, p = 0.026, \eta^2 = 0.07$). As Fig. 2.10a illustrates, when outcomes were collisions, witnessing AVs executing illegal discretionary actions reduced general trust in AVs. This pattern appears to have reversed when outcomes were near-misses. However, neither of these differences reached statistical significance ($p = 0.129, p = 0.097$, respectively).

Ratings of likelihood of adoption were also analyzed with a mixed ANOVA that revealed a non-significant main effect of Stage in Experiment 1 ($F(1, 63) = 0.37, p = 0.545$). However there was a significant Stage X Outcome interaction ($F(1, 63) = 4.63, p = 0.035, \eta^2 = 0.07$). As Fig. 2.9b illustrates, witnessing legal discretionary actions that led to collisions does not have a significant impact on future likelihood of adopting AVs ($p = 0.283$), whereas there is a marginally non-significant positive impact when these actions led to near-misses ($p = 0.054$). In contrast, there was a significant main effect of Stage in Experiment 2 ($F(1, 74) = 7.00, p = 0.010, \eta^2 = 0.09$). Overall, scenarios featuring AVs executing illegal discretionary actions seem to have lowered participants' likelihood of AV adoption ($M = 3.76, SE = 0.35$ pre-trial versus $M = 3.24, SE = 0.35$ post-trial). However, as in Experiment 1, there was a significant Stage X Outcome interaction ($F(1, 74) = 5.14, p = 0.026, \eta^2 = 0.07$). As Fig. 2.10b illustrates, the likelihood of adoption was only significantly lowered for the Outcome: Accident condition ($p < 0.001$), not the Outcome: Near-miss condition ($p = 0.784$).

Overall, the measures of general acceptance of AVs reveal that legal discretionary actions do not have detrimental impacts on people's trust in AVs, regardless of outcomes. Witnessing them can even potentially improve general acceptance of AVs when the outcome is positive. However, AVs performing illegal discretionary actions could potentially lower people's trust in the technology in general but only when these actions led to collisions – the trust level maintains if these actions led to near-misses.

[3] As with blame, the potential effect of individual differences in driving experience on trust ratings was explored in both experiments using a 2 (Driving Experience: driver, non-driver) X 2 (Operator: autonomous system, human driver) X 2 (Outcome: collision, near-miss) X 2 (Situation Acuteness: high, low) mixed ANOVA. There was no significant effect of driving experience on trust in either experiment ($F(1, 121) = 0.27, p = 0.602$; $F(1, 144) = 0.23, p = 0.631$) or significant interaction between driving experience and any other independent variable ($ps > 0.05$), except for with Situation Acuteness in Experiment 2 ($F(1, 144) = 6.04$, $p = 0.015, \eta^2 = 0.04$), which was not the primary focus of this Experiment or paper). Thus, driving experience was excluded from further analyses.

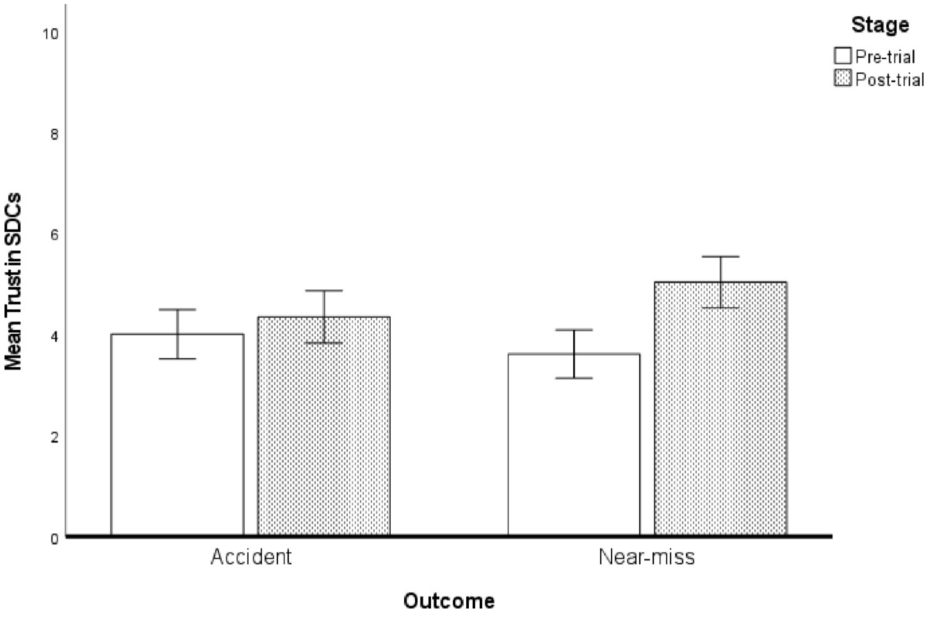


Fig. 2.9a. Mean ratings of general trust in AVs before and after being exposed to scenarios across both outcome conditions in Experiment 1 (Error bars = ±1 SE).

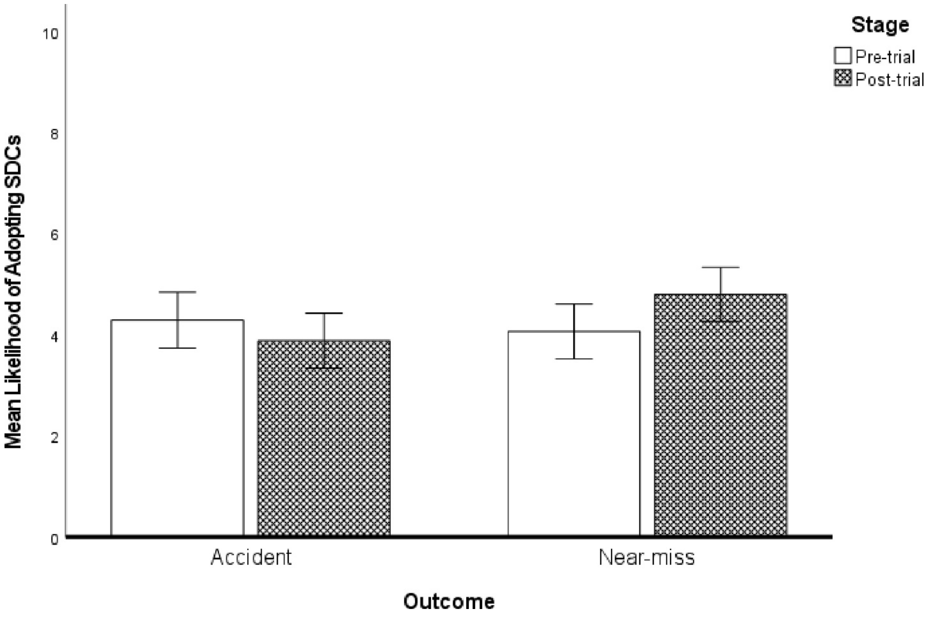


Fig. 2.9b. Mean ratings of likelihood of adopting AVs before and after being exposed to scenarios across both outcome conditions in Experiment 1 (Error bars = ±1 SE).

Taken together, the results of Experiment 1 and 2 suggest that public acceptance of AVs performing legal discretionary action is fairly high, but acceptance of illegal discretionary actions of AVs is lower, both compared to legal actions and human drivers performing the same actions. Also, there seems to be a fundamental shift in the judgemental frame when discretionary actions become illegal, from one that focuses on *how well* the action was executed when it was legal, to *whether* it should have been executed when it was illegal. However, based on these findings, can we safely infer that, the public's acceptance of AVs' "inaction" (i.e., not perform illegal actions) would be higher than human drivers in the same circumstance, or that people would prefer AVs to *not* act in these circumstances? Experiment 3 was designed to address these questions: to what extent would an AV versus HD be blamed and trusted when opting *not* to perform illegal discretionary actions leading to negative or positive outcomes.

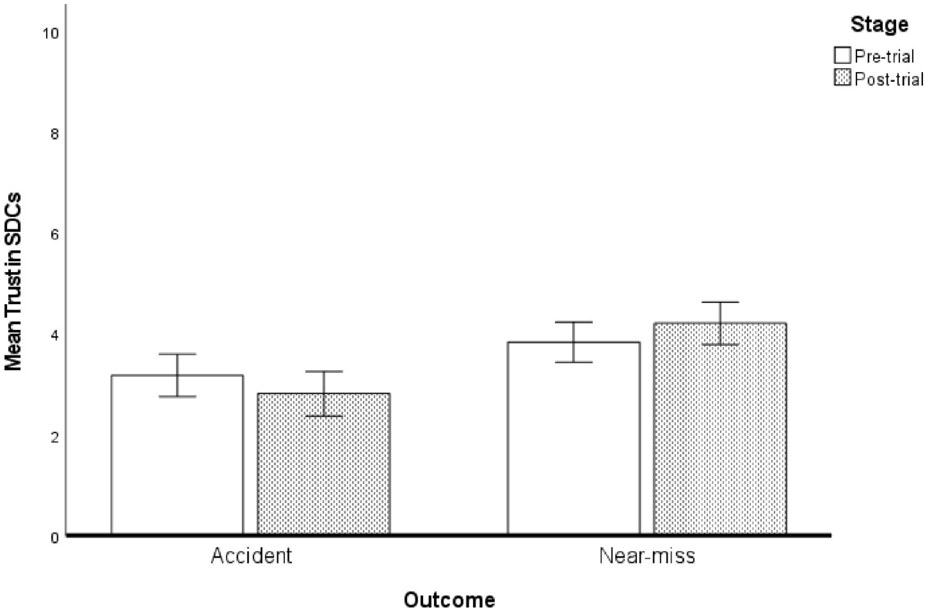


Fig. 2.10a. Mean ratings of general trust in AVs before and after being exposed to scenarios across both outcome conditions in Experiment 2 (Error bars = ±1 SE).

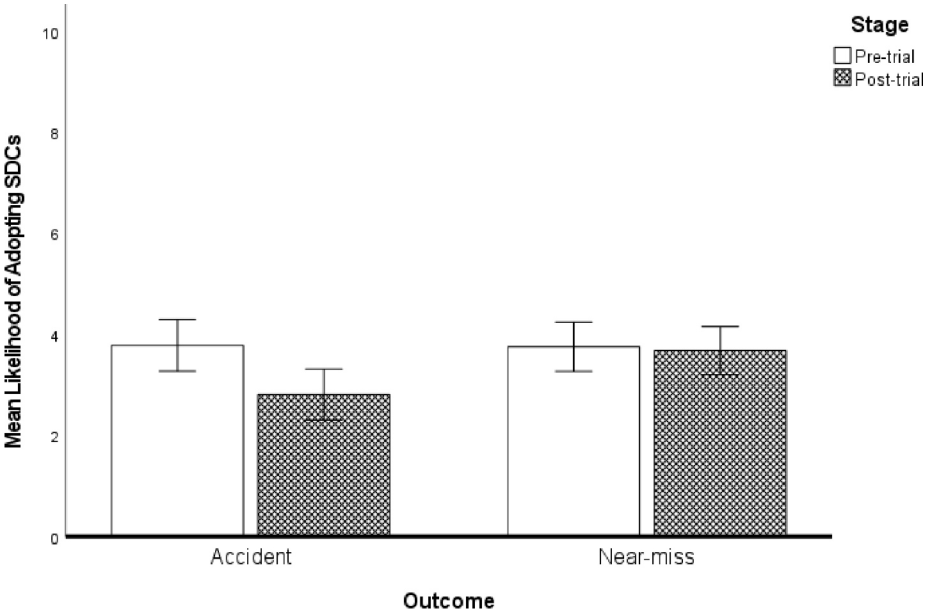


Fig. 2.10b. Mean ratings of likelihood of adopting AVs before and after being exposed to scenarios across both outcome conditions in Experiment 2 (Error bars = ±1 SE).

6. Experiment 3

The focus of Experiments 1 and 2 has been on discretionary action. However, we must also consider the opposite: omission of a discretionary action – or inaction. The results of Experiment 2 suggest that judgements of AVs are harsher when they perform an illegal discretionary action 1) compared to a legal action (Experiment 1) and 2) compared to a human driver performing the same action (Experiments 1 and 2). Does this mean that people would prefer highly autonomous vehicles to not perform discretionary actions in the same situations (that is – preference for inaction)? Experiment 3 was designed to explore this question, featuring scenarios where AVs or HD vehicles chose not to perform illegal discretionary actions, leading to either a positive (i.e., improving traffic flow or narrowly avoiding a collision) or negative outcome (i.e., causing congestion or having a collision). It was predicted that blame and trust ratings would be reciprocal to those of Experiments 1 and 2 if we assume that human preference systems are cohesive. More specifically, blame on AVs should be lower than that on human drivers when they choose not to perform the same illegal action in the same circumstance leading to the same outcome.

6.1. *Method*

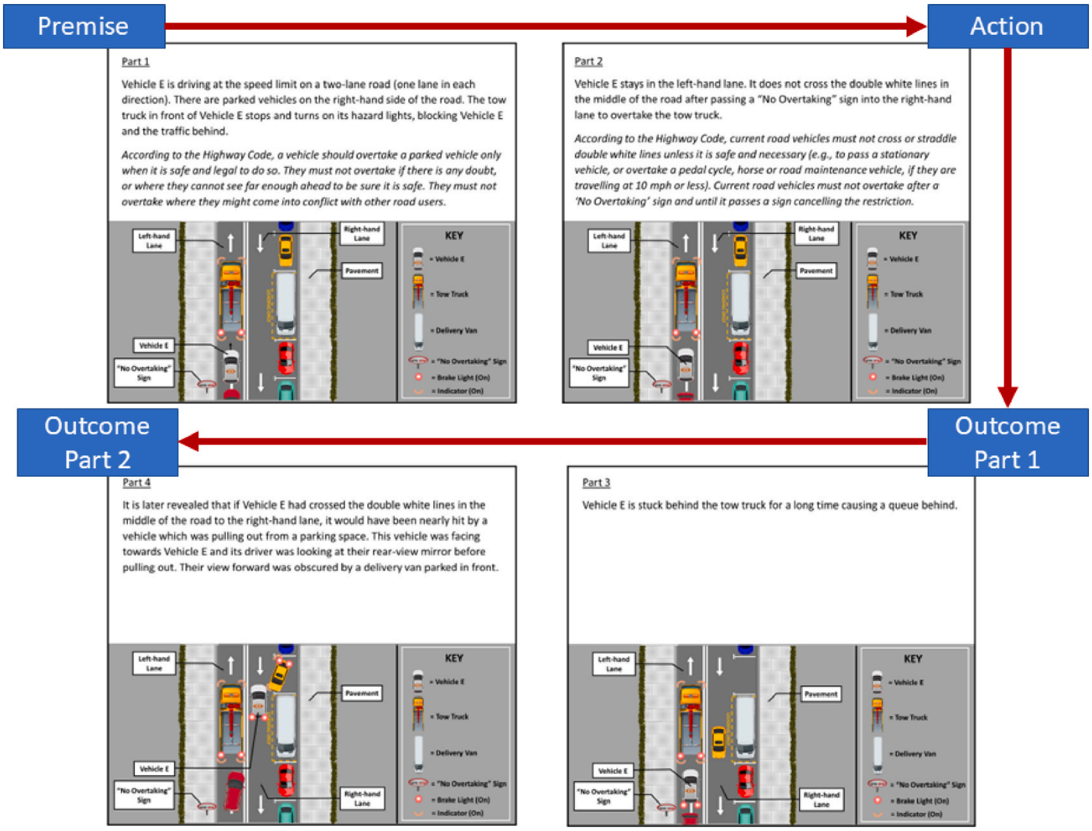
6.1.1. *Participants*

One-hundred and seventy-one participants were recruited using the same method and inclusion and exclusion criteria as in Experiments 1 and 2. The sample consisted of 83 females (48.5%) and 87 males (50.9%) and one non-disclosed individual. The mean age was 37.80 (*SD* = 13.39, *Min* = 18, *Max* = 78). One-hundred-and-twenty-five participants (73.1%) had a full driving license. Of those, the average number of years of driving experience was 20.92 (*SD* = 15.96, *Min* = 1, *Max* = 61) and the average annual mileage of driving was 6561.78 (*SD* = 5711.08, *Min* = 0, *Max* = 40,000).

6.1.2. *Design, materials and procedure*

This experiment featured adapted versions of the nine driving scenarios from Experiment 2. Each consisted of four parts instead of three (see Fig. 3.1 – Full Scenarios can be found in Appendix D). Part 1 was the same as in Experiment 2. Part 2 described the decision that was made by the target vehicle: instead of committing an illegal discretionary action (i.e., straddling the double white line in the middle of the road), it was described as having stayed in its lane and braked (in the Acuteness: High scenarios) or waited (in the Acuteness: Low scenarios). Incident outcomes were delivered during the last two parts: Part 3 described the "factual" aspect of the outcome (e.g., Vehicle A does not stop in time and hits the pedestrian) whereas Part 4 described the "counterfactual" aspect (e.g., If Vehicle A had crossed the double white lines in the middle of the road to the right-hand lane, it would have avoided the pedestrian but would have been nearly hit by a vehicle, which was backing out of a driveway). In the previous two experiments, these two aspects were delivered in one part (i.e., Part 3). The reason they were displayed separately in the current experiment was because the counterfactual element was much more difficult to visualize than in Experiment 1 & 2 and an additional pictorial illustration was required.

IVs and DVs were identical to the previous experiments, as was the procedure. A summary of participants' allocation to each condition can be found in Table 1.2. It should however be noted that although the two versions of the outcomes in this experiment can



Part 1

Vehicle E is driving at the speed limit on a two-lane road (one lane in each direction). There are parked vehicles on the right-hand side of the road. The tow truck in front of Vehicle E stops and turns on its hazard lights, blocking Vehicle E and the traffic behind.

According to the Highway Code, a vehicle should overtake a parked vehicle only when it is safe and legal to do so. They must not overtake if there is any doubt, or where they cannot see far enough ahead to be sure it is safe. They must not overtake where they might come into conflict with other road users.

Part 2

Vehicle E stays in the left-hand lane. It does not cross the double white lines in the middle of the road after passing a "No Overtaking" sign into the right-hand lane to overtake the tow truck.

According to the Highway Code, current road vehicles must not cross or straddle double white lines unless it is safe and necessary (e.g., to pass a stationary vehicle, or overtake a pedal cycle, horse or road maintenance vehicle, if they are travelling at 10 mph or less). Current road vehicles must not overtake after a 'No Overtaking' sign and until it passes a sign cancelling the restriction.

Part 3

Vehicle E is stuck behind the tow truck for a long time causing a queue behind.

Part 4

It is later revealed that if Vehicle E had crossed the double white lines in the middle of the road to the right-hand lane, it would have been nearly hit by a vehicle which was pulling out from a parking space. This vehicle was facing towards Vehicle E and its driver was looking at their rear-view mirror before pulling out. Their view forward was obscured by a delivery van parked in front.

Fig. 3.1. Structure and flow of the nine vignettes. *Note.* the example features a scenario in which the target vehicle needs to negotiate a parked tow truck in front and ended with a negative outcome.

roughly be mapped onto the previous two (i.e., similar in valence), they were not exactly comparable. The negative outcomes in this experiment entailed collisions as well as consequences that were far less serious (e.g., hindered traffic flow). Similarly for positive outcomes, this experiment featured near-misses as well as mundane situations (e.g., improved traffic flow). In contrast, Experiment 1 & 2 only featured collisions and near-misses.

6.2. *Results and discussion*

The correct rate in responses to the nine attention questions was 100%, indicating all participants acted diligently and attentively when undertaking the experiment. For the sake of consistency and comparability, the nine scenarios were grouped into two categories in the same way in Experiment 1 and 2, based on their acuteness.

6.2.1. *Blame*

Figs. 3.2a and 3.2b illustrates mean ratings of blame on the various parties involved across all conditions. Note that the scenarios in the current experiment only involved one third-party (equivalent to Third Party A in previous experiments). In general, blame levels on the target vehicle were mostly lower than that on the third party in all cases, especially in Acuteness High scenarios. The following analyses will focus on blame on the target vehicle, which is of the primary interest of the current paper.

There was a significant main effect of Outcome ($F(1, 167) = 70.22$, $p < 0.001$, $\eta^2 = 0.30$) on level of blame on the target vehicle. Unsurprisingly, blame was markedly higher when the outcome was negative ($M = 3.54$, $SE = 0.17$) than positive ($M = 1.52$, $SE = 0.17$) (see Fig. 3.3). However, the magnitude of this difference seems to be dependent on the acuteness of the situation, indicated by a significant Outcome X Situation Acuteness interaction ($F(1, 167) = 15.49$, $p < 0.001$, $\eta^2 = 0.09$). As Fig. 3.3 shows, as the situation became less acute, the difference in blame between the two outcome conditions reduced ($ps < 0.001$). This seemed to occur because the degree of blame after negative outcomes was reduced as the acuteness of the situation decreased ($p < 0.001$), but blame level following a positive outcome did not ($p = 0.125$). This pattern was very similar to that of Experiment 2.

There was also a significant main effect of Operator ($F(1, 167) = 6.45$, $p = 0.012$, $\eta^2 = 0.04$). Contrary to the prediction, blame was generally higher when the vehicle was operated by an autonomous system ($M = 3.86$, $SE = 0.17$) than a human ($M = 2.83$, $SE = 0.21$) for not performing the discretionary action (see Figs. 3.2a and 3.2b). This was consistent across scenarios of different situation acuteness, supported by a non-significant Operator X Situation Acuteness interaction ($F(1, 167) = 0.47$, $p = 0.496$). All other main effects and interactions were non-significant ($p > 0.05$).

Compared to Experiment 2 that featured *AVs committing* illegal discretionary actions, ratings of blame on AVs *omitting* to perform discretionary actions seem to be relatively lower in similar circumstances, indicating a preference for inaction. However, these results should be interpreted with caution because the scenarios used in the two experiments were not directly comparable. For example, outcomes of “action” scenarios in Experiment 2 featured collisions with minor injuries or near-misses, whereas outcomes of “inaction” (Experiment 3) were more diverse and some much less sever (e.g., causing congestion). Interestingly, blame ratings do not display a reciprocal pattern to what was observed in Experiment 2 in terms of autonomous systems versus human drivers’ comparisons. Instead, they display a similar pattern where autonomous systems were blamed more than human drivers regardless of whether they performed or opted not to perform an illegal discretionary action. Hence, on the surface, it seems to suggest that participants did not have a clear preference over whether illegal actions should be committed by an AV or not. Instead, the consequence of that action or inaction seems to be the dominating factor in their blame judgement.

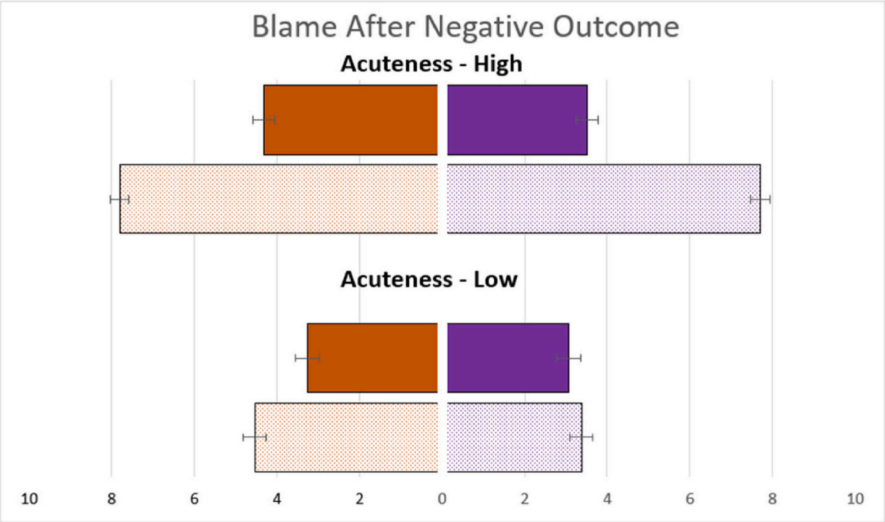


Fig. 3.2a. Mean ratings of blame on the target vehicle and third parties in the Outcome: negative condition (Error bars = ±1 SE).

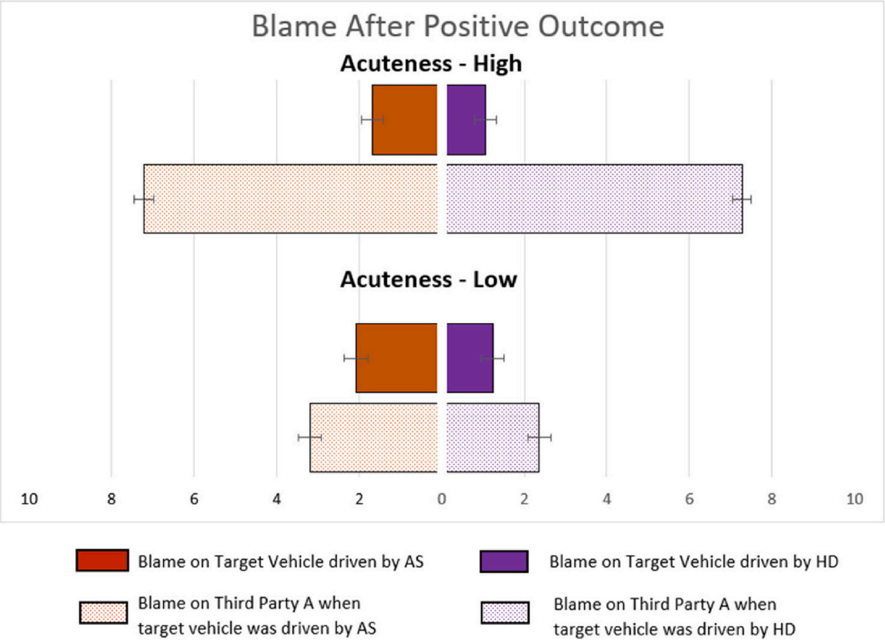


Fig. 3.2b. Mean ratings of blame on the target vehicle and third parties in the Outcome: positive condition (Error bars = ±1 SE).

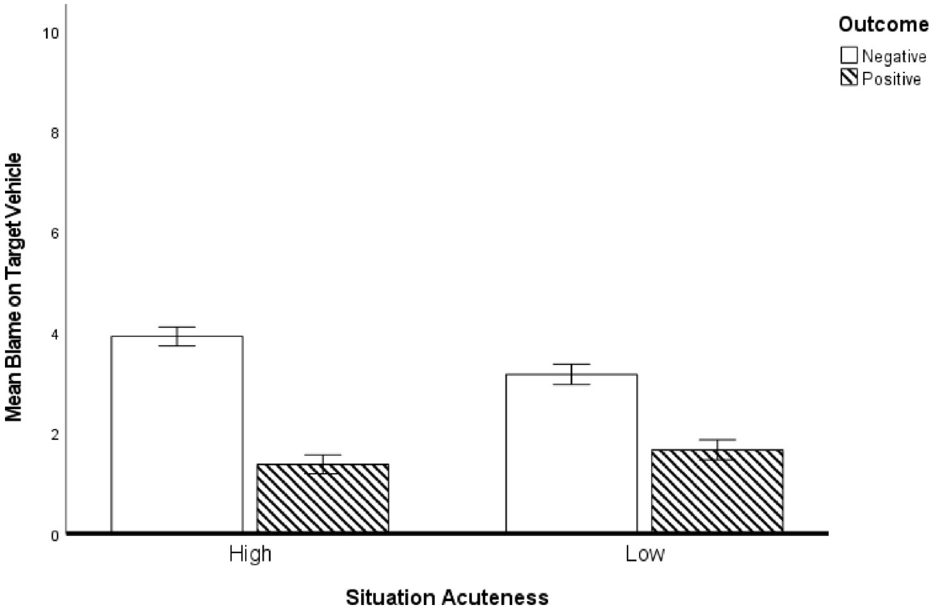


Fig. 3.3. Mean ratings of blame on the target vehicle in scenarios of different situation acuteness across two outcome conditions (Error bars = ±1 SE).

6.2.2. Operator-specific trust

As Fig. 3.4 illustrates, there was a significant main effect of Situation Acuteness ($F(1, 167) = 24.39$, $p < 0.001$, $\eta^2 = 0.13$) and Outcome ($F(1, 167) = 41.31$, $p < 0.001$, $\eta^2 = 0.20$) on ratings of trust in the driver/operator of the target vehicle. There was also a significant interaction ($F(1, 167) = 39.70$, $p < 0.001$, $\eta^2 = 0.19$). When situation acuteness was high, a positive outcome induced a higher level of trust than when the outcome was negative ($p < 0.001$). This difference seems to have been reduced (but was significant) when situation acuteness was low ($p < 0.001$), mainly due to the fact that only the trust ratings after negative outcomes increased as situation acuteness decreased ($p < 0.001$) whereas ratings after positive outcomes were relatively stable ($p = 0.336$). There was also a significant main effect of Operator ($F(1, 167) = 45.47$, $p < 0.001$, $\eta^2 = 0.21$). Human drivers were trusted trust ($M = 7.40$, $SE = 0.19$)

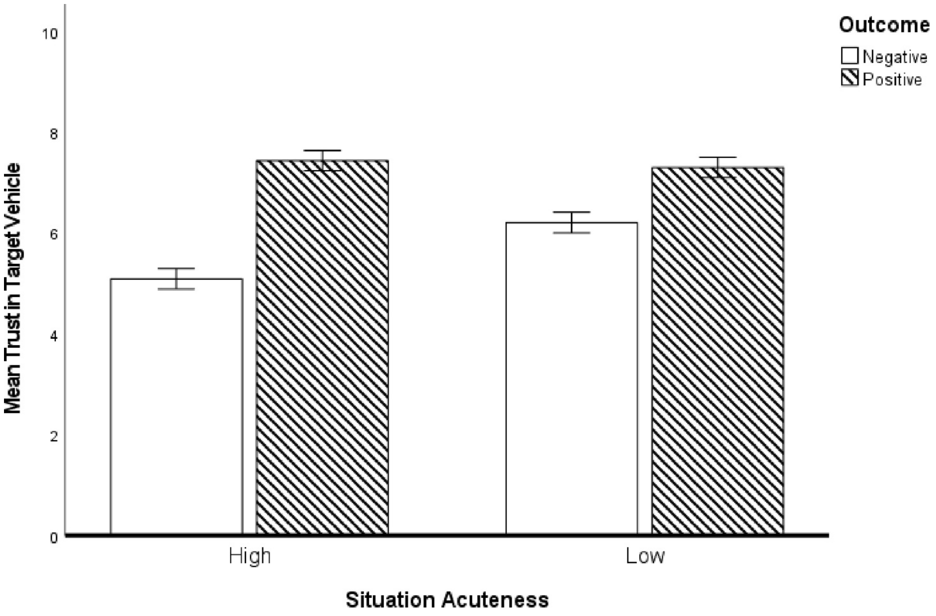


Fig. 3.4. Mean ratings of trust in the target vehicle in scenarios of different situation acuteness across two outcome conditions (Error bars = ±1 SE).

than autonomous systems ($M = 5.60$, $SE = 0.19$). This is consistent between Acuteness High and Low scenarios, evidenced by a lack of significant interaction between Situation Acuteness and Operator ($F(1, 167) = 0.56$, $p = 0.456$, $\eta^2 < 0.01$). All other main effects and interactions were non-significant ($p > 0.05$).

Overall, trust findings in specific vehicle operators display a similar pattern to that of the previous two experiments. There was a significant negative correlation between operator-specific trust and blame overall but the strength of correlation was stronger in the HD (r = -0.72, $p < 0.001$) than AS condition (r = -0.49, $p < 0.001$). For the AS condition in particular, operator-specific trust was also strongly correlated with pre-existing acceptance (i.e., general trust in and likelihood of adopting AVs) (both $rs = 0.44$, $p < 0.001$) whereas there were no significant correlations between blame and pre-existing acceptance ($r = 0.19$, $p = 0.095$; $r = 0.14$, $p = 0.221$, respectively). These findings are consistent with that of the previous experiments and reinforce the notion that apar from blame, operator-specific trust is also influenced by pre-existing acceptance of AVs in general.

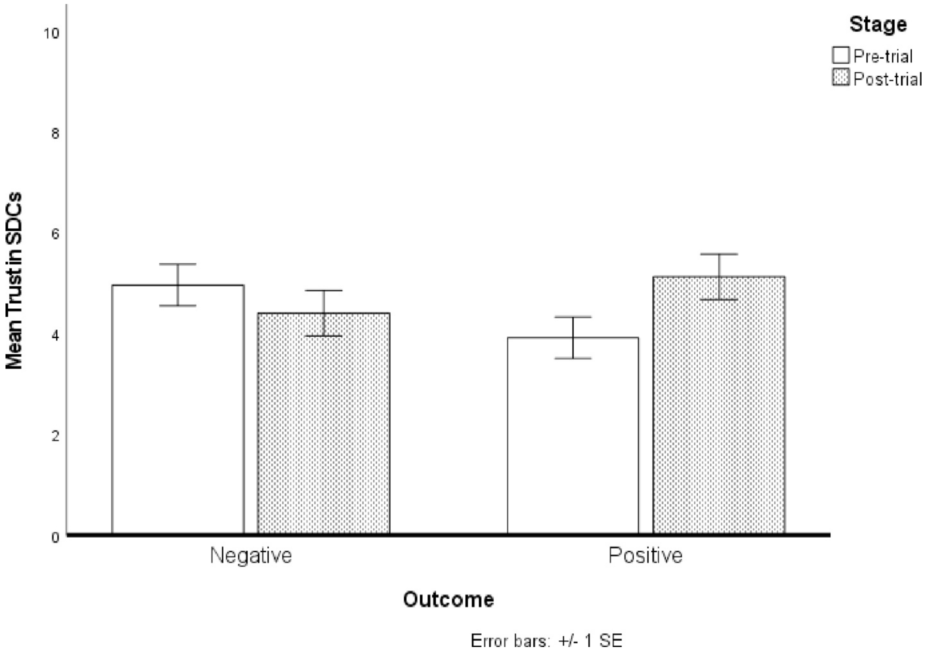


Fig. 3.5. Mean ratings of general trust in AVs before and after being exposed to scenarios across both outcome conditions (Error bars = ±1 SE).

6.2.3. General trust towards AVs and likelihood of adoption

As in Experiments 1 and 2, participants' general acceptance of AVs in the AS condition were measured both before and after reading the nine scenarios. There was a non-significant main effect of Stage ($F(1, 84) = 2.31$, $p = 0.132$) and Outcome ($F(1, 84) = 0.08$, $p = 0.776$) on general trust and a significant interaction between the two variables ($F(1, 84) = 17.01$, $p < 0.001$, $\eta^2 = 0.17$). As Fig. 3.5 illustrates, when outcomes were positive, witnessing AVs *not* executing illegal discretionary actions during the main phase of the Experiment increased trust in them ($p < 0.001$). This pattern seems to have reversed when outcomes were negative, although the difference was not significant ($p = 0.069$).

The ratings of likelihood of adoption were also analyzed and displayed a similar pattern as trust. There was a non-significant main effect of Stage ($F(1, 83) = 0.14$, $p = 0.713$) and Outcome ($F(1, 83) = 0.04$, $p = 0.847$) though both variables significantly interacted ($F(1, 83) = 15.86$, $p < 0.001$, $\eta^2 = 0.16$). As Fig. 3.6 illustrates, when the outcome was negative, witnessing an AV not executing a discretionary manoeuvre decreased the likelihood of adoption ($p = 0.003$). The effect was reversed when the outcome was positive ($p = 0.013$). Overall, these results reflect a very outcome-oriented nature of the general attitude towards AVs. That is, participants' attitude change seem to have been primarily influenced by the valence of the consequence of an action or inaction, rather than the consideration of whether the action should or should not have been committed.

7. General discussion

In order to gain public trust, acceptance and possible adoption, autonomous vehicles (AVs) not only need to operate safely in a complex traffic environment but also behave in a manner that is conducive to the smooth operation of the transport network(s) as well as conforms with human user expectations. We stress that this would require some level of flexibility and adaptability to be built into their behavioural specifications which allow for improvisation and discretion, instead of such technology having to consistently follow pre-set, hard-coded traffic rules without any scope for adaptability, regardless of often dynamically changing circumstances. Such flexibility appears not only intuitive but necessary given the complexities of road transport networks and users (e.g., other AVs, drivers of non-AVs, pedestrians, cyclists, and so on), where strict traffic rules and laws cannot always be adhered to (e.g., in the event of an emergency or indeed in discretionary situations where traffic flow is not optimal but could be). However, a key issue is whether and to what extent the public would accept additional risks – i.e. AVs stretching the boundaries of the ROTR / Highway Code − that will inevitably come with an even higher level of autonomy.

In the experiments presented in the current paper, participants' judgements of blame and trust were taken after reading vignette stories featuring AVs performing (or not performing) legal (Experiment 1) or illegal (Experiments 2 and 3) discretionary actions in response to traffic system conditions with varying degrees of acuteness, which ranged from e.g., overtaking a cyclist (low acuteness) to e.g., swerving to avoid an inevitable collision with a pedestrian walking into the road (high acuteness). Within our scenarios, we also added a subsequent situation in which, having performed the manoeuvre, the AV found itself in a situation where a negative outcome either occurred (e.g., colliding with another road user) or almost occurred (e.g., a near miss with another road user). Overall, the findings revealed a promising but complex picture of attitudes (specifically for trust and blame) towards AVs executing (Experiments 1 and 2) or choosing not to execute (Experiment 3) discretionary actions.

First, after AVs performed legal discretionary actions leading to a collision or a near-miss (Experiment 1), blame on AVs was relatively low, both in terms of the adopted rating scale (all means were below the midpoint of the scale) as well as compared to other parties involved in the same incidents (i.e., parties that induced the discretionary actions and those whose actions directly preceded the collisions/near-misses). This was especially true when the situation acuteness was high (i.e., the operator of the vehicle faced an emergency and needed to react quickly). Also, notably the level of blame on AVs in Experiment 1 was in general on par with that of Zhang et al. (2024) – the scenarios of which mainly featured high acuteness situations – and did not specify a discretionary action as a choice for AVs. This indicates that the explicit knowledge that an AV has committed a legal (according to the ROTR / Highway Code) discretionary action did not drastically affect the judgement of blame, implying a reasonably high level of acceptance of these actions. However, blame on an AV was markedly higher in the event of performing illegal discretionary actions (Experiment 2) even when those actions led to the same unforeseen negative outcomes as in Experiment 1 (i.e., collisions). Nonetheless, when illegal discretionary actions (Experiment 2) led to more positive outcomes (i.e., near-misses), the level and the pattern of blame was very akin to when the action was legal (Experiment 1). That is, blame on AVs was mostly lower than that of the other parties. Overall, it appears that the acceptance of illegal discretionary actions is lower than that of legal ones, which is not a complete surprise given that illegal actions represent a more severe level of rule violation. But one interesting aspect of this finding is that outcome valence (i.e. near miss vs collision) appears to be a more dominant force affecting attitudes than the legality of discretionary actions, pointing to an "ends-justify-the-means" mentality. This point will be discussed in more detail in a later section.

Second, our hypothesis that AVs should be blamed more than human drivers after performing discretionary actions was only partly supported. Although on most occasions, AVs were found to be blamed significantly more than human drivers for performing the same manoeuvres under the same situations, the magnitude of these inter-operator differences was arguably small from a practical point of view (effect sizes measured via η^2 being smaller than 0.04) and also inconsistent. That is, when the action was legal (Experiment 1), AVs were blamed more than human drivers only when situation acuteness was high, whereas blame was not significantly different between AVs and human drivers when it was low. In contrast, when the discretionary action was illegal (Experiment 2), AVs were blamed uniformly more than human drivers, regardless of the situation acuteness. Hence, our hypothesis seems to have been only fully supported in situations where AVs performed an illegal discretionary action. It should be noted that our hypothesis regarding AVs-human driver parity was based on the assumption that AVs performing discretionary actions would violate people's perceptions that they are rule-based systems and hence prompt counterfactual thoughts (e.g., "They shouldn't have done so") more readily than for

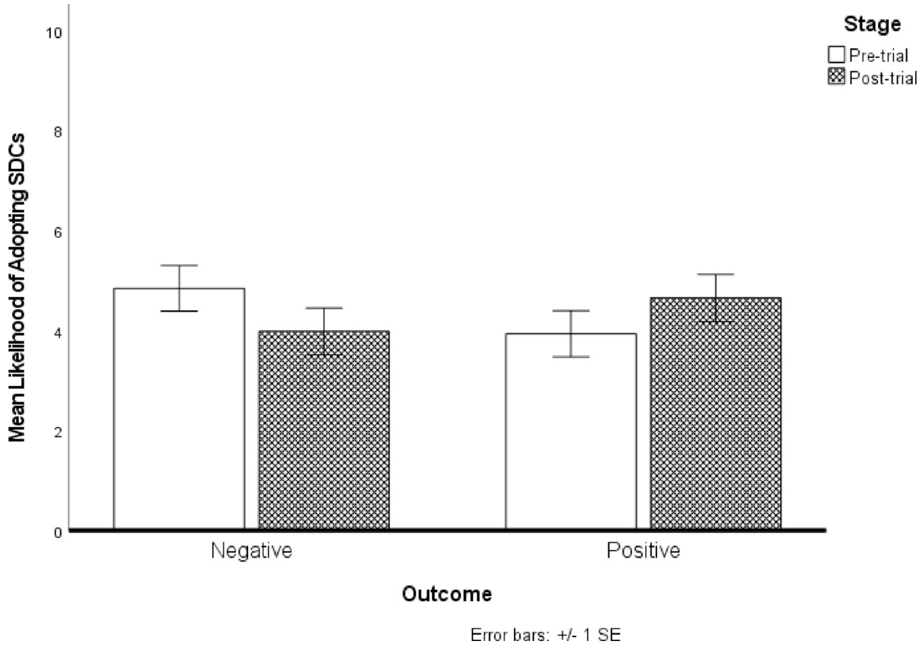


Fig. 3.6. Mean ratings of likelihood of adopting AVs before and after being exposed to scenarios across both outcome conditions (Error bars = ±1 SE).

human drivers (Kahneman & Miller, 1986; Markman & McMullen, 2003; Miller et al., 1990; Roese & Olson, 1997). The fact that the relative levels of blame on AVs versus human drivers displayed different patterns depending on the legality of the discretionary actions points to the possibility that the proposed psychological mechanism does not apply to all situations and that different mental processes might be involved depending on the legality of the actions and the acuteness of the situation.

One possible way that the mental process may have shifted, is that people's judgements in the two experiments anchored at different aspects of the incident depending on whether the action was legal or illegal. More specifically, when the discretionary actions in question were legal, the focus of judgement seems to be on "performance" – in other words, *how well* a manoeuvre was executed. In these circumstances, participants' rating of blame displayed a pattern consistent with the *Adaptive Capability Perception Hypothesis* (Zhang et al., 2024). That is, AVs are perceived to be superior compared to human drivers in terms of reaction times and accuracy and hence should be blamed more when failing to prevent an accident in acute situations. In contrast, human drivers are perceived to be superior to AVs in terms of improvisation and inductive causal reasoning and hence should be blamed more when failing to prevent an accident in non-acute situations that require discretion and judgements. On the other hand, when discretionary actions were illegal, the focus of judgement seems to have shifted to the justifiability of the action itself, i.e. *whether* it should have been executed in the first place. Under these circumstances, the actions of the AV violated the stereotypical perception that machines are rule-based systems and hence they were less excusable than human drivers for breaking the law even when they did so to resolve an emergency situation.

Third, a reliable difference in operator-specific trust was observed between AVs and human drivers in every experiment. That is, an AV was in general trusted less than a human driver after it performed a discretionary action. However, this difference could be simply due to the fact that AVs were trusted less in general to begin with (Yokoi, 2024), rather than because their specific actions in the scenarios were judged less favourably than human drivers. Supporting this possibility, correlational analyses revealed that the relationship between blame and operator-specific trust was weaker for AVs than for human drivers. Also, operator-specific trust in AVs was strongly correlated with baseline pre-test acceptance of self-driving technology in general. These correlational patterns also echo those of Zhang et al. (2024) suggesting that for AVs, trust in a specific operator is heavily influenced by trust in the AV technology in general, whereas for conventional vehicles, trust in a particular human driver is mostly based on the performance of that specific driver.

The higher reliance on the attitudes towards the "group" to derive the opinion of a "member" could be reflective of stereotypical perceptions that AVs capable of SAE L4 mode operation will likely soon be mass-produced and deployed at scale and hence are perceived as homogenous in their quality and characteristics. By comparison, people might perceive that human drivers are different from each other at least in some respect and hence need to be evaluated according to their own merits. This sense of homogeneity of AVs could also be caused by the fact that people find AI and robots, although having achieved a high level of autonomy, unfamiliar and alien to them. Therefore, judgements are subject to what is called the "out-group homogeneity" effect (Park & Rothbart, 1982; Rubin & Badea, 2012), which is characterised by people's tendency to perceive social groups that are outside of their own as less diverse and differentiated. Regardless of the causes, this sense of homogeneity has important practical implications. While an unpleasant encounter with one *suboptimal* human driver can often be quickly dismissed as an isolated event (e.g., a bad apple), the experience of suboptimal performance/behaviour of one AV is more likely to be generalised as a trait of the entire group (Quattrone & Jones, 1980). It is a machine and thus likely to be 'programmed' to perform that suboptimal action over and over again. This highlights the fragility of people's trust in this technology especially at a stage where they are not very familiar with it, yet are the potential future user group who will inevitably determine the success of this disruptive technological leap within the road transport domain. Undesirable behaviours, especially those leading to negative traffic incidents (such as collisions), can be quickly translated into doubt and loss of confidence in the technology as a whole. This strengthens our arguments on the importance of optimally specifying AVs behaviour characteristics during the crucial period of early introduction – they 'should' perform in general far more optimally than human drivers although they will never be flawless.

Reinforcing those points, the findings show that observing AVs performing legal discretionary actions could potentially improve human 'general' trust in self-driving technology as well as the likelihood of acceptance and possible adoption when those actions (e.g., to potentially avoid a collision) lead to outcomes that are not highly negative (i.e., near-misses) but still involve an element of risk. When the discretionary actions in question were illegal, there was evidence that these could have detrimental impacts on general acceptance (e.g., trust) but only when the actions led to collisions. When they resulted in near-misses, both general trust and likelihood of adoption appear to have sustained their original levels pre-exposure to scenarios. These results are quite remarkable in that 1) they show how susceptible people's general acceptance of self-driving technology is to the knowledge of traffic incidents, despite the scenarios being highly abstract and the participants being explicitly informed of their hypothetical and contrived nature; and 2) people's reactions to near-misses appear to have been quite positive, despite the fact that by definition, near-misses are situations in which things could have easily gone wrong and hence are often taken as signs of danger or existing defects in a system (Teigen, 1998, 2005; Tinsley et al., 2012). It was perhaps surprising to find that – in terms of trust –participants seem to have shown a high level of acceptance of AVs performing manoeuvres that led to these dangerous situations even when these manoeuvres were illegal.

Near-misses can provide conflicting messages about the state of a system (Dillon & Tinsley, 2016; McMullen & Markman, 2002; Zhang & Covey, 2014) – the positive message that a *disaster* has been successfully avoided (and hence the system is 'still safe') and the negative message that a disaster could have taken place (and hence the system is 'still safe' but 'dangerous'). It appears that participants focused more on the positive message of the near-misses instead of the possible negative outcomes (though future experiments are needed to e.g., compare varying degrees of near miss situations). This could be especially true in the high acuteness condition where the danger was brought about by a third party and the vehicle seemed forced to take action. In these scenarios, the dangerous situations were more likely to be accepted as part of the "background" of the story and the vehicles' behaviours, on the other hand, are perceived as the "force" to correct or resolve the danger. In comparison, in the low situation acuteness scenarios, there was no immediate danger to the vehicle and its action was hence more likely to be treated as a factor contributing to the danger and therefore the near-misses would appear more "self-inflicted". This line of reasoning received some support from the fact that in both Experiment 1 and 2, blame after a near-miss was generally higher (and operator-specific trust lower) if situation acuteness was low rather than high. However, participants in our experiments were not offered a chance to rate their general acceptance or comment on the two types of scenarios separately, leaving room for future research.

To further probe participants' attitudes towards discretionary actions, their responses to AVs' "inactions" were investigated in Experiment 3. When the AV chose not to execute illegal discretionary actions, blame was lower (and trust higher) than when committing such actions (as in Experiment 2). This complements findings from Experiment 2 and suggests a public preference towards "inaction" (i.e. the AV should come to a stop – perhaps even irrespective of the outcome of that action). However, it should be noted that although Experiment 3 employed materials and design that were largely the same as those in Experiment 2, outcomes of some scenarios had to be changed to aid plausibility. More specifically, the negative outcomes of some scenarios were more benign (e.g., causing traffic congestion) than those in Experiment 2 (e.g., collision). Therefore, the results of Experiments 2 and 3 were not entirely comparable. Furthermore, the pattern of blame and trust ratings from Experiment 3 was not completely reciprocal to that of Experiment 2, indicating a lack of internal coherency in people's judgemental framework. For example, AVs were found to be blamed more than human drivers in both experiments, whether they committed or did not commit illegal discretionary actions. AVs' inactions also led to a lowered likelihood of adoption when these inactions resulted in negative outcomes, just like when AVs committed those actions in Experiment 2. These findings seem to be sending mixed signals with respect to whether or not AVs should execute discretionary actions in a given situation because either way they seem to have lowered people's acceptance as long as the outcomes were negative, indicating a judgemental system that's dominated by the consideration of outcome valence instead of AVs' choices.

Although the effect of outcome valence was not the main point of interest, it was the most reliable predictor of post-incident judgements, consistently producing effects on blame and operator-specific trust ratings in all three experiments. It was also found to be an important factor in determining general acceptance of self-driving technologies. This is not surprising considering that negative emotions have been found to be a major contributing factor in the attribution of blame (Feigenson & Park, 2006; Malle et al., 2013). Collisions, especially those involving personal injuries, are more likely to provoke emotional reactions than less severe outcomes like near-misses and hence were understandably judged more harshly. However, it can be argued that it is irrational to judge the merits of an action based on its consequences. Outcome bias refers to the tendency for humans to evaluate the quality of a decision based on its outcome instead of its intention, rationale and information available to the decision maker at the time (Baron & Hershey, 1988). In our scenarios, whether the discretionary actions led to collisions or near-misses was determined by factors outside of the control of the vehicle operator, but they were somehow found to have a huge bearings on people's judgement of blame and trust. The prevalence of this "ends-justify-the-means" type of mentality can pose great challenges to vehicle manufacturers and policy makers since no behavioural protocols can guarantee optimal outcome in every situation. The nature and extent of the outcome bias should be investigated further by future research in the context of post-collision judgements involving AVs. An example of experimental design would be to measure preferences and acceptance of an AV's decision before the outcome is revealed and compare them to those measured after the outcome is known.

8. Practical implications

Above and beyond the generally promising picture, our results revealed great complexity in people's judgement process regarding their attitudes towards AVs performing discretionary actions. Thus, we need to take a more nuanced and cautionary approach when it comes to policy making and legislation regarding the discretionary actions that AVs could or indeed should be able to perform. First, the fact that participants were more critical towards AVs' illegal actions compared to legal actions (possibly due to stereotypical perception of computers being rule-based machines) as well as the fact that they applied seemingly distinctive judgemental frameworks to the two situations, suggests the necessity to conceptually distinguish these two types of actions both in regulation/law making and system programming. For example, the parameters to satisfy a potential future condition that warrants an AV to perform an illegal discretionary action (in response to e.g., emergency situations) should be stricter than for legal actions, both in terms of necessity (e.g., not performing the action unless absolutely necessary) and risk tolerance (e.g., not performing the action unless extremely safe). Second, the findings highlight several features of people's evaluation process that might distort judgements towards AV behaviours. For example, the perception of product homogeneity might make people more prone to over-generalisation of the undesirable behaviour of one AV to the entire fleet or even the technology. This would make the isolated negative incidents particularly impactful.

Furthermore, the tendency to overweight outcome valence in evaluating AVs' decision quality (i.e., outcome bias), although understandable, could distract people from scrutinising the context and rationale of AVs' decision-making. An AV's decision could be judged unfavourably just because it causes a collision despite it being caused by factors beyond the AV's control or knowledge. This could also lead to situations where AV manufacturers have to modify decision protocols after negative incidents due to public pressure which does not necessarily lead to better decision quality. This highlights the need to establish formal principles and criteria for decision appraisals and educate the public on their rationales. This is particularly important given that AVs are not expected to eliminate road collisions and accidents involving AVs are thought to be inevitable (Fagnant & Kockelman, 2015).

9. Limitations and future directions

There are limitations within and across the current experiments. Although the findings indicate more critical attitudes towards illegal, as opposed to legal AV manoeuvres, it should be noted that the "law" in this context refers to the current ROTR prescribed for human drivers. Their high relevance is justified by the fact that many regulatory bodies around the world (e.g., Automated Vehicles Act, 2024; NHTSA, 2017; United Nations Economic Commission for Europe, 2020) have advocated strong compliance with the local traffic rules (for humans) when specifying AV driving behaviours. However, the status-quo might change in the future when the presence of AVs increases, and the public are more accustomed to using and/or sharing roads with them. That is, a different set of rules might be written specifically for AVs than for human drivers, taking into account their capabilities and characteristics. For example, Koopman et al (2019) proposed the idea of *Digital Highway Code* (DHC) which does not only include ROTR but also specifications for good AV behaviours under exceptional circumstances. Hence, behaviours like straddling a solid white line, mounting on the kerb or even speeding might well be within the boundary of laws for AVs. Although legalizing such behaviours would be practically similar to allowing AVs to engage in illegal manoeuvres at their discretion, it might provoke a drastically different psychological experience and attitudes, which remains to be tested by future studies.

Many interesting inferences were drawn from our findings relating to public attitudes towards legal, as opposed to illegal discretionary actions of AVs as well as towards actions versus inactions. However, many of these inferences were derived from comparisons of findings *between* experiments, instead of between conditions *within* experiments using statistical tests. Despite our best effort to match conditions under which the data was collected; experiments might differ in aspects other than the variables being manipulated. For example, Experiment 1 and 2 were conducted at slightly different temporal points and the characteristics of their samples also differed in terms of gender and age compositions as well as average driving experience. Although there is no reason to believe that one or more of these variables would have systematically affected the results (noting that negative news stories could have some influence), caution needs to be taken when interpreting the differences observed between the two experiments with respect to attitudes towards legal versus illegal manoeuvres of AVs. Future research should set out to test those inferences in settings where the legality of discretionary actions is formally manipulated.

Being entirely novel and partly exploratory in nature, our experiments were not designed in such a way to test the full range of aforementioned speculations. Some of the key psychological constructs invoked by the above discussions, including participants' preferences in the operator's choices (i.e., action versus inaction), and the assumed psychological mechanism of counterfactual thinking, were not directly measured but instead inferred from the patterns of other dependent variables including blame ratings. Future research can begin to test these by attempting to directly measure participants' preferences and the content of their counterfactual thinking process through, for example, spontaneous thought-generating tasks (e.g. Meyers-Levy & Maheswaran, 1992; Roese & Olson, 1997). Furthermore, to manage the total duration of each experiment and to prevent fatigue, many constructs, including operator-specific trust and blame measured after each scenario, as well as the general trust in AVs and intention to use, were only crudely measured by a single question. While this practice is not uncommon in the research of AVs (for examples, see Liu & Du, 2021; Pöllänen et al., 2020; Zhang et al., 2024) and has the advantage of capturing the fleeting, holistic mental state of the respondents, it is in part limited in representing the multidimensional and complex nature of those constructs. The results of such measures also do not allow for checks of internal consistency and reliability and hence are subject to higher likelihood of measurement errors. Future studies should employ more sophisticated measurement methods – e.g., multi-item scales (e.g., Malle & Ullman, 2021) as well as objective measures such as those that can track physiological variations (e.g., eye gaze tracking – Hergeth et al., 2016)) to achieve a higher level of reliability and validity.

10. Conclusions

It might feel ironic to some that in order to gain the trust of the public and to ensure the functioning of the traffic system, AVs, at least at their early stage of deployment, need to possess some driving characteristics of human drivers. However, equipping AVs with the ability to exercise discretion and deviate from hard traffic rules might come at the price of heightened risk of subsequent accidents, which, although likely infrequent, would also occur if a human driver were in control, and importantly were not foreseeable before the initial discretionary action was committed to. While the results of our experiments indicate that the public seems to be willing to extend a similar level of tolerance and leniency towards AVs as towards human drivers, they also revealed features of human judgements that might pose challenges to the deployment of a flexible system. The understanding of these features, both on the part of the vehicle manufacturers and policy makers, will be vital to the proliferation of self-driving technology. The experiments described here represent an initial step towards this direction. However, with a narrow range of contrived scenarios and crude measurements, they have only scratched the surface of human perceptions of AV behaviours and left much room for further investigations into the mental process underpinning attitude forming and their characteristics, including cognitive biases.

Author Note

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 – ES/T007079/1 (Morgan, UK Principal Investigator; Inatani, Japan Principal Investigator).

CRediT authorship contribution statement

Qiyuan Zhang: Writing – review & editing, Writing – original draft, Visualization, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Victoria Marcinkiewicz:** Methodology, Investigation, Conceptualization. **Louise Bowen:** Methodology, Investigation, Conceptualization. **Theodor Kozlowski:** Resources, Methodology, Conceptualization. **Tatsuhiko Inatani:** Methodology, Funding acquisition, Conceptualization. **Yoshiyuki Ueda:** Methodology, Conceptualization. **Hirofumi Katsuno:** Methodology, Funding acquisition, Conceptualization. **Minoru Asada:** Methodology, Funding acquisition, Conceptualization. **Phillip L. Morgan:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.tra.2026.104885.

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. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, 897–902. 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: Pol. Pract., 118, 819–836. 10.1016/J.TRA.2018.10.019.

Ahmed, H.U., Huang, Y., Lu, P., Bridgelall, R., 2022. Technology developments and impacts of connected and autonomous vehicles: an overview. Smart Cities 5 (1), 382–404. https://doi.org/10.3390/smartcities5010022.

Automated Vehicles Act, 2024. UK Government Bill. https://www.legislation.gov.uk/ukpga/2024/10/contents.

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.

Baron, J., Hershey, J., 1988. Outcome bias in decision evaluation. J. Pers. Soc. Psychol. 54, 569–579. https://doi.org/10.1037//0022-3514.54.4.569.

Bellem, H., Schönenberg, T., Krems, J.F., Schrauf, M., 2016. Objective metrics of comfort: developing a driving style for highly automated vehicles. Transport. Res. F: Traffic Psychol. Behav. 41, 45–54. https://doi.org/10.1016/j.trf.2016.05.005.

Bellem, H., Thiel, B., Schrauf, M., Krems, J.F., 2018. Comfort in automated driving: an analysis of preferences for different automated driving styles and their dependence on personality traits. Transport. Res. F: Traffic Psychol. Behav. 55, 90–100. https://doi.org/10.1016/j.trf.2018.02.036.

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.

Bin-Nun, A.Y., Derler, P., Mehdipour, N., Tebbens, R.D., 2022. How should autonomous vehicles drive? Policy, methodological, and social considerations for designing a driver. Human. Soc. Sci. Commun. 9 (1), 299. https://doi.org/10.1057/s41599-022-01286-2.

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.

Bonnefon, J.F., Shariff, A., Rahwan, I., 2019. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars. Proc. IEEE 107 (3), 502–504. https://doi.org/10.1109/JPROC.2019.2897447.

Brunello, A., Montanari, A., Reynolds, M., 2019. Synthesis of LTL formulas from natural language texts: State of the art and research directions. In: 26th International Symposium on Temporal Representation and Reasoning (TIME 2019).

Chakraborty, D., Chaisse, J., Pahari, S., 2020. Global auto industry and product standards: a critical review of India's economic and regulatory experience. J. Int. Trade Law Pol. 19 (1), 8–35.

Chen, Y., Gandhi, R., Zhang, Y., Fan, C., 2023. Nl2tl: Transforming natural languages to temporal logics using large language models. ArXiv Preprint ArXiv: 2305.07766.

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

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.

Dillon, R.L., Tinsley, C.H., 2016. Near-miss events, risk messages, and decision making. Environment Systems and Decisions 36, 34–44. https://api.semanticscholar.org/CorpusID:146776609.

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.

Edelmann, A., Stümper, S., Petzoldt, T., 2021. Cross-cultural differences in the acceptance of decisions of automated vehicles. Appl. Ergon. 92, 103346. https://doi.org/10.1016/j.apergo.2020.103346.

Endicott, T., 2001. Law is necessarily vague. Leg. Theory 7 (4), 379–385. https://doi.org/10.1017/S135232520170403X.

Endicott, T., 2011. Vagueness and Law. In G. Ronzitti (Ed.), Vagueness: A Guide (pp. 171–191). Springer Netherlands. 10.1007/978-94-007-0375-9_7.

Fagnant, D., Kockelman, K., 2015. Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transp. Res. A Policy Pract. 77. https://doi.org/10.1016/j.tra.2015.04.003.

Feigenson, N., Park, J., 2006. Emotions and attributions of legal responsibility and blame: a research review. Law Hum Behav. 30 (2), 143–161. https://doi.org/10.1007/s10979-006-9026-z.

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: P. M. Fitts (Ed.), Human engineering for an effective air-navigation and traffic-control system. National Research Council, Div. of.

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., Billings, D.R., Schaefer, K.E., Chen, J.Y.C., de Visser, E.J., Parasuraman, R., 2011. A meta-analysis of factors affecting trust in human-robot interaction. Hum. Factors 53 (5), 517–527. https://doi.org/10.1177/0018720811417254.

Hancock, P.A., Nourbakhsh, I., Stewart, J., 2019. On the future of transportation in an era of automated and autonomous vehicles. Proc. Natl. Acad Sci. - 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. 10.1080/00140139.2018.1441448, 61(8), 1017–1032. 10.1080/00140139.2018.1441448.

Helson, H., 1964. Adaptation-level theory: an experimental and systematic approach to behavior. In Adaptation-level theory: an experimental and systematic approach to behavior. New York.

Hergeth, S., Lorenz, L., Vilimek, R., Krems, J.F., 2016. Keep your scanners peeled: gaze behavior as a measure of automation trust during highly automated driving. Hum. Factors 58 (3), 509–519. https://doi.org/10.1177/0018720815625744.

Highway Code, Highway Code, 2023.

HM Government, 2022. Connected & Automated Mobility 2025: Realising the benefits of self-driving vehicles in the UK.

Hoff, K.A., Bashir, M., 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. Hum. Factors 57 (3), 407–434.

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

Hong, J.W., Cruz, I., Williams, D., 2021. AI, you can drive my car: how we evaluate human drivers vs. self-driving cars. Comput. Hum. Behav. 125, 106944. https://doi.org/10.1016/J.CHB.2021.106944.

Institution of Mechanical Engineers, 2023. Public Perceptions: Autonomous Vehicles - Survey Results.

International Organization for Standardization, 2025. Road vehicles — Safety for automated driving systems — Design, verification and validation (ISO/TS 5083: 2025). https://www.iso.org/standard/81920.html.

Jelinski, L., Etzrodt, K., Engesser, S., 2021. Undifferentiated optimism and scandalized accidents: the media coverage of autonomous driving in Germany. J. Sci. Commun. 20 (4), A02.

Kahneman, D., Frederick, S., 2002. Representativeness revisited: attribute substitution in intuitive judgment. Heuristics 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.

Koopman, P., Hierons, R., Khastgir, S., Clark, J., Fisher, M., Alexander, R., Eder, K., Thomas, P., Barrett, G., Torr, P., 2019. Certification of highly automated vehicles for use on UK roads: Creating an industry-wide framework for safety.

Lazarus, R.S., 1991. Emotion and adaptation. In Emotion and adaptation. Oxford University Press.

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

Lee, J., Moray, N., 1992. Trust, control strategies and allocation of function in human-machine systems. Ergonomics 35 (10), 1243–1270. https://doi.org/10.1080/00140139208967392.

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

Liu, P., Du, Y., Xu, Z., 2019a. Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. Accident Anal. Prevent. 125, 232–240. https://doi.org/10.1016/J.AAP.2019.02.012.

Liu, P., Yang, R., Xu, Z., 2019b. How safe is safe enough for self-driving vehicles? Risk Anal.. 39 (2), 315–325. https://doi.org/10.1111/risa.13116.

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

Makridis, M.A., Anesiadou, A., Mattas, K., Fontaras, G., Ciuffo, B., 2023. Characterising driver heterogeneity within stochastic traffic simulation. TransportmetricaB: Transp. Dyn. 11 (1), 725–743. https://doi.org/10.1080/21680566.2022.2125458.

Malle, B.F., Guglielmo, S., Monroe, A.E., 2013. Moral, cognitive, and social: The nature of blame. In Social thinking and interpersonal behavior. (pp. 313–331). Psychology Press.

Malle, B. F., Ullman, D., 2021. Chapter 1 - A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), Trust in Human-Robot Interaction (pp. 3–25). Academic Press. 10.1016/B978-0-12-819472-0.00001-0.

Markman, K.D., Gavanski, I., Sherman, S.J., McMullen, M.N., 1993. The mental simulation of better and worse possible worlds. J. Exp. Soc. Psychol. 29 (1), 87–109. https://doi.org/10.1006/jesp.1993.1005.

Markman, K.D., McMullen, M.N., 2003. A reflection and evaluation model of comparative thinking. Personal. Soc. Psychol. Rev. 7 (3), 244–267. https://doi.org/10.1207/S15327957PSPR0703_04.

Mayer, R.C., Davis, J.H., Schoorman, F.D., 1995. An integrative model of organizational trust. Acad. Manage. Rev. 20 (3), 709–734. https://doi.org/10.2307/258792.

McMullen, M.N., Markman, K.D., 2002. Affective impact of close counterfactuals: implications of possible futures for possible pasts. J. Exp. Soc. Psychol. 38 (1), 64–70. https://doi.org/10.1006/jesp.2001.1482.

Meyers-Levy, J., Maheswaran, D., 1992. When timing matters: the influence of temporal distance on consumers' affective and persuasive responses. J. Consum. Res. 19 (3), 424–433. https://doi.org/10.1086/209312.

Miller, D.T., Turnbull, W., McFarland, C., 1990. Counterfactual Thinking and Social Perception: Thinking about What Might Have Been. In M. P. Zanna (Ed.), Advances in Experimental Social Psychology (Vol. 23, pp. 305–331). Academic Press. 10.1016/S0065-2601(08)60322-6.

Moody, J., Bailey, N., Zhao, J., 2020. Public perceptions of autonomous vehicle safety: an international comparison. Saf. Sci. 121, 634–650. https://doi.org/10.1016/j.ssci.2019.07.022.

Naiseh, M., Clark, J., Akarsu, T., Hanoch, Y., Brito, M., Wald, M., Webster, T., Shukla, P., 2025. Trust, risk perception, and intention to use autonomous vehicles: an interdisciplinary bibliometric review. AI & Soc. 40 (2), 1091–1111. https://doi.org/10.1007/s00146-024-01895-2.

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, 120319. https://doi.org/10.1016/j.techfore.2020.120319.

Nees, M.A., 2016. Acceptance of self-driving cars: an examination of idealized versus realistic portrayals with a self- driving car acceptance scale. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 60 (1), 1449–1453. https://doi.org/10.1177/1541931213601332.

NHTSA, 2017. Automated Driving Systems 2.0: A Vision for Safety.

Nyhan, R., 2000. Changing the paradigm: trust and its role in public sector organizations. Am. Rev. Public Administrat. - AMER REV PUBLIC ADM 30, 87–109. https://doi.org/10.1177/02750740022064560.

Olaverri-Monreal, C., 2020. Promoting trust in self-driving vehicles. Nat. Electron. 3 (6), 292–294. https://doi.org/10.1038/s41928-020-0434-8.

Othman, K., 2022. Exploring the implications of autonomous vehicles: a comprehensive review. Innovat. Infrastruct. Solut. 7 (2), 165. https://doi.org/10.1007/s41062-022-00763-6.

Parasuraman, R., Riley, V., 1997a. Humans and automation: use, misuse, disuse. Abuse. Human Factors 39 (2), 230–253. https://doi.org/10.1518/001872097778543886.

Parasuraman, R., Riley, V., 1997b. Humans and automation: use, misuse, disuse, abuse. Hum. Factors. https://doi.org/10.1518/001872097778543886.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cognit. Eng. Decis. Mak. 2 (2), 140–160. https://doi.org/10.1518/155534308X284417.

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.

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.

Prakken, H., 2017. On the problem of making autonomous vehicles conform to traffic law. Artificial Intelligence and Law 25 (3), 341–363. https://doi.org/10.1007/s10506-017-9210-0.

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.

Rahman, Md. M., Thill, J.-C., 2024. Who is inclined to buy an autonomous vehicle? Empirical evidence from California. Transportation. 10.1007/s11116-024-10490-9.

Robertson, C.E., Pröllochs, N., Schwarzenegger, K., Pärnamets, P., Van Bavel, J.J., Feuerriegel, S., 2023. Negativity drives online news consumption. Nat. Hum. Behav. 7 (5), 812–822. https://doi.org/10.1038/s41562-023-01538-4.

Roese, N.J., 1994. The functional basis of counterfactual thinking. J. Pers. Soc. Psychol. 66 (5), 805–818. https://doi.org/10.1037/0022-3514.66.5.805.

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. Curr. Dir. Psychol. Sci. 21 (6), 367–372. https://doi.org/10.1177/0963721412457363.

SAE, 2021. SAE Levels of Driving Automation™ Refined for Clarity and International Audience.

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., Hancock, P.A., 2016. A meta-analysis of factors influencing the development of trust in automation: implications for understanding autonomy in future systems. Hum. Factors 58 (3), 377–400. https://doi.org/10.1177/0018720816634228.

Schmitt, A., 2020. Right of Way: Race, Class, and the Silent Epidemic of Pedestrian Deaths in America. Island Press.

Schneble, C.O., Shaw, D.M., 2021. Driver's views on driverless vehicles: public perspectives on defining and using autonomous cars. Transp. Res. Interdiscip. Perspect. 11, 100446. https://doi.org/10.1016/j.trip.2021.100446.

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 Y. Shionoya & K. Yagi (Eds.), Competition, Trust, and Cooperation (pp. 3–19). Springer Berlin Heidelberg.

Smith, B.W., 2017. How governments can promote automated driving. NML Rev. 47, 99.

Teigen, K.H., 1998. When the unreal is more likely than the real: Post hoc probability judgments and counterfactual closeness. Think. Reason. 4 (2), 147–177. https://doi.org/10.1080/135467898394193.

Teigen, K.H., 2005. The proximity heuristic in judgments of accident probabilities. Br. J. Psychol. 96 (4), 423–440. https://doi.org/10.1348/000712605X47431.

Tennant, C., Neels, C., Parkhurst, G., Jones, P., Mirza, S., Stilgoe, J., 2021. Code, culture, and concrete: self-driving vehicles and the rules of the road. Front. Sustainable Cities 3. https://doi.org/10.3389/frsc.2021.710478.

Tennant, C., Stares, S., Howard, S., 2019. Public discomfort at the prospect of autonomous vehicles: building on previous surveys to measure attitudes in 11 countries. Transport. Res. F: Traffic Psychol. Behav. 64, 98–118. https://doi.org/10.1016/j.trf.2019.04.017.

Tinsley, C.H., Dillon, R.L., Cronin, M.A., 2012. How near-miss events amplify or attenuate risky decision making. Manag. Sci. 58 (9), 1596–1613. https://doi.org/10.1287/mnsc.1120.1517.

Türkoğlu, İ.K., Bilici, F., 2024. Autonomous vehicle technology and technology acceptance: the role of technological readiness on consumers&#039; attitudes towards driverless cars and intention to use in the Future. Fırat Üniversitesi Mühendislik Bilimleri Dergisi 36 (1), 383–407. https://doi.org/10.35234/fumbd.1385541.

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.

United Nations Economic Commission for Europe, 2020. Proposal for a New UN Regulation on Uniform Provisions Concerning the Approval of Vehicles with Regards to Automated Lane Keeping System. UNECE.

Wallbridge, C., Zhang, Q., Marcinkiewicz, V., Bowen, L., Kozlowski, T., Jones, D., Morgan, P., 2024. "Warning!" benefits and pitfalls of anthropomorphising autonomous vehicle informational assistants in the case of an accident. Multimodal Technol. Interact. 8, 110. https://doi.org/10.3390/mti8120110.

Watson, J., van der Linden, S., Watson, M., Stillwell, D., 2024. Negative online news articles are shared more to social media. Sci. Rep. 14 (1), 21592. https://doi.org/10.1038/s41598-024-71263-z.

Wu, J., Liao, H., Wang, J.-W., 2020. Analysis of consumer attitudes towards autonomous, connected, and electric vehicles: a survey in China. Res. Transp. Econ. 80, 100828. https://doi.org/10.1016/j.retrec.2020.100828.

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. Transp. Res. Part C Emerging Technol. 95, 320–334. https://doi.org/10.1016/j.trc.2018.07.024.

Yokoi, R., 2024. Trust in self-driving vehicles is lower than in human drivers when both drive almost perfectly. Transport. Res. F: Traffic Psychol. Behav. 103, 1–17. https://doi.org/10.1016/j.trf.2024.03.019.

Zhang, Q., Covey, J., 2014. Past and future implications of near-misses and their emotional consequences. Exp. Psychol. 61 (2), 118–126. https://doi.org/10.1027/1618-3169/a000231.

Zhang, Q., Wallbridge, C.D., Jones, D.M., Morgan, P.L., 2024. Public perception of autonomous vehicle capability determines judgment of blame and trust in road traffic accidents. Transp. Res. A Policy Pract. 179, 103887. https://doi.org/10.1016/j.tra.2023.103887.

Zhang, Q., Wallbridge, C.D., Jones, D., Morgan, P., 2021. The blame game: double standards apply to autonomous vehicle accidents. Lecture Notes in Networ. Syst. 270, 308–314. https://doi.org/10.1007/978-3-030-80012-3_36.