Received: 19 December 2023 | Accepted: 2 January 2025

DOI: 10.1002/jcpy.1448

RESEARCH ARTICLE

# Public perception and autonomous vehicle liability

**Julian De Freitas**[1] | **Xilin Zhou**[2] | **Margherita Atzei**[2] | **Shoshana Boardman**[3] | **Luigi Di Lillo**[4]

[1]Marketing Unit, Harvard Business School, Cambridge, Massachusetts, USA

[2]Property and Casualty Solutions, Reinsurance, Swiss Reinsurance Company, Ltd, Zürich, Switzerland

[3]University of Oxford, Oxford, UK

[4]Zardini Lab, MIT, Cambridge, Massachusetts, USA

**Correspondence**
Julian De Freitas, Marketing Unit, Harvard Business School, Cambridge, MA, USA.
Email: jdefreitas@hbs.edu

**Abstract**

The deployment of autonomous vehicles (AVs) and the accompanying societal and economic benefits will greatly depend on how much liability AV firms will have to carry for accidents involving these vehicles, which in turn impacts their insurability and associated insurance premiums. Across three experiments ($N$=2677), we investigate whether accidents where the AV was not at fault could become an unexpected liability risk for AV firms, by exploring consumer perceptions of AV liability. We find that when such accidents occur, the not-at-fault vehicle becomes more salient to consumers when it is an AV. As a result, consumers are more likely to view as relevant counterfactuals in which the not-at-fault vehicle might have behaved differently to avoid or minimize damage from, the accident. This leads them to judge AV firms as more liable than both firms that make human-driven vehicles and human drivers for damages when not at fault.

**KEYWORDS**

autonomous vehicles, harm, insurance, liability, moral judgment

## INTRODUCTION

Every year globally, around 1.25 million people are killed by motor vehicle accidents on our roads and 20 million more are injured. Human error—and the systems that make it easy for these errors to be dangerous (Nader, 1965; Welle et al., 2018)—is responsible for 90% of these accidents (Singh, 2015).

Fully autonomous vehicles (AVs), which perform driving tasks without human intervention or assistance, promise to improve this status quo. Societally, AVs could obey speed limits and are incapable of getting distracted, tired, stressed, angry, or drunk (Koopman & Wagner, 2017). They could reduce congestion by driving optimally; free human attention to converse, conduct meetings, drive tired, or just sleep; provide more freedom for those who struggle to get transport; and, because most AVs will use electric or hybrid drivetrains rather than internal combustion engines, they could reduce our carbon footprint on the planet. Economically, AVs could enable shuttle and ride-sharing firms to offer their services 24/7, without worker capacity limits or the costs of employing human drivers. Of course, whether AVs truly end up delivering on these various benefits is a complex issue and requires consideration of important downsides and unintended consequences triggered by this technology (De Freitas et al., 2022).

Despite these promises, as the public starts to encounter and use AVs that function in increasingly broad operating conditions—as is already happening in Austin, Las Vegas, Phoenix, and San Francisco (Carlson, 2022; Kolodny, 2022; Randazzo, 2020; Wessling, 2022)—accidents are inevitable, increasing the liability risk for AV firms. In the United States, for example, makers of driver assistance technologies (a lower level of automation than fully autonomous vehicles) have already faced a stream of accident-related lawsuits for issues such as defective steering sensors and camera misalignments (Smith, 2017, 2022; Villasenor, 2014). Most notably, Cruise, a subsidiary of General Motors, recently lost its license to operate in San Fransisco after one of its autonomous vehicles was involved in an accident where it was not at fault (Bensinger, 2024; Cano, 2024).

Here, we approach this question by exploring how consumers perceive the liability of AV manufacturers in the common scenario where the AV is not at fault.

Accepted by Jennifer Argo and David B. Wooten, Editors; Associate Editor, Zachary Estes.

  

While not-at-fault accidents are not typically considered liability risks for human drivers or manufacturers of human-driven vehicles (HDVs), we ask whether the public thinks the firm that manufactured the not-at-fault vehicle is more liable when the vehicle was an AV than HDV, posing a liability risk for AV manufacturers and an existential threat to AV adoption. In what follows, we present the conceptual background and our theoretical framework, followed by three studies that test the proposed response pattern and relevance of the counterfactual thinking process. These results are further buttressed by four supplemental studies, reported in the Appendix S1. We conclude with theoretical and practical implications.

## CONCEPTUAL BACKGROUND

### Product liability for autonomous vehicles

Businesses are vulnerable to lawsuits when they are causally connected to defects in their offerings (Loudenback & Goebel, 1974; Morgan, 1987). In fact, firms may be held liable even if they abided by existing regulations in the production and sale of their offerings, because they are ultimately judged on whether they behaved "unreasonably" by not taking alternative actions to prevent harm. Given this standard, a significant challenge for firms is anticipating all the scenarios in which they could be judged as unreasonable—even potential "edge cases" like a consumer using their product in unlikely ways, for example, driving a tire at exceedingly high speeds.

Thus far, the automotive industry has not been liable for most motor vehicle accidents. If two regular human drivers (HDVs) are involved in a motor vehicle accident, they have the choice to settle through a traditional insurance policy or to engage in litigation against the vehicle manufacturer (HG.org, 2024). Since most such vehicle accidents result from driver error, they fall under the legal banner of "driver negligence," such that the at-fault driver (or their insurance) pays the damages. If the accident results from some defect in the vehicle itself, however, it falls under "product liability," a form of commercial liability in which a firm or its insurance covers the damages instead. Such product liability cases make up only 6% of regular HDV motor vehicle-related claims (Smith, 2017).

An open question is how liability risks play out when, inevitably and increasingly, AVs are involved in motor accidents. Since AV firms make the AI-aided software stack responsible for the driving task, if an AV is *at fault* then "driver error" should now be the firm's responsibility, and is expected to fall under commercial liability, in particular the traditional banner of product liability (Smith, 2017).

Here, we consider the less intuitive possibility that AV firms will be held liable even when they are involved in accidents where they are *not* at fault because their vehicles will be viewed as defective. We get at this question by measuring ascriptions of liability by consumers. Consumers are pertinent for several reasons. First, lawsuits against AV firms, which are most likely to go to trial when victims are seriously injured, will involve consumers as plaintiffs. In these cases, the awards will be economically significant for firms. For instance, the median plaintiff verdict in cases involving HDVs can range from $5 million (in the event of victim death) to $14 million (quadriplegia) (Smith, 2017), and it balloons for class actions lawsuits on behalf of a larger group (https://www.law.cornell.edu/wex/class_action). Even if only some cases reach a jury, precedent suggests that the results of these trials will set the benchmark for settlements that take place outside of court (Smith, 2017).

Second, consumers will make up the juries that decide how much to award in these cases. Aspects of juror psychology, such as juror sympathy for the defendant, may affect product liability awards (Darden et al., 1991).

Third, the liability judgments of consumers can be an alternative measure of liability risk in the absence of claims and other driving data. Since AVs are not yet widely deployed, there is a dearth of liability claims data available (Wells, 2022), making it difficult for insurers and risk managers to apply traditional approaches to estimate the liability risks of AV firms (SwissRe, 2022). Additionally, AVs have not yet driven sufficient miles to afford a statistically meaningful failure rate comparison (injuries and fatalities) with HDVs (Kalra & Paddock, 2016). As a result, manufacturers and their insurers must turn to alternative approaches and risk measures to study, estimate, explain, and ultimately take on the potential liability risks of AVs.

Finally, the AV industry has not yet adequately articulated a concept of AV defectiveness (Smith, 2017), which will need to cover not just the hardware but also the software responsible for the driving task. In the absence of formal definitions of AV defectiveness, public perception biases can have a greater impact.

### AI failures

Perhaps the closest related work to ours is on how consumers respond to cases in which AI fails when it is *at fault*. Most of this work finds an effect that goes in the opposite direction to the one we are predicting here: AI is viewed as *less* blameworthy than humans for the same error. For example, in the domain of autonomous vehicles, human drivers are blamed more than their automated cars when both make mistakes, in partially

automated settings where a human may take over control of the vehicle or vice versa (Awad et al., 2019). This appears to be because, compared to humans, AI is viewed as being less agentic and intentional than human decision makers (Arikan et al., 2023; Li et al., 2016; Srinivasan & Sarial-Abi, 2021).

## Autonomous vehicle adoption

Also related is work on the adoption of AVs, which finds that, despite the economic and societal advantages of AVs, consumers prefer to avoid riding in them. For instance, 63% of consumers say they would not want to ride in an AV if given the opportunity (Rainie et al., 2022), 76% feel less safe riding in cars with self-driving features, and 79% would not pay more to own a car with self-driving features (Brennan & Sachon, 2022). Many hesitations stem from safety concerns over the performance and failures of autonomous vehicles, as well as fear of ceding control to a machine (Schoettle & Sivak, 2014; Shariff et al., 2021), and concerns about how AVs will make difficult moral tradeoffs like whether to crash into a group of elderly pedestrians or swerve into a barrier and thereby kill the passengers it contains (De Freitas et al., 2020, 2021; De Freitas & Cikara, 2021).

Much of this work finds that AV adoption boils down to whether consumers trust AVs enough to ride in them (Gold et al., 2015; Xu et al., 2018), where trust is typically defined as a willingness to become vulnerable with another because one has positive expectations about them (Rousseau et al., 1998, p. 395). In the context of AV adoption, trust can be treated as a willingness to make oneself vulnerable to an autonomously behaving product whose operation is outside of one's own control. Consumer vulnerability in this context is clear because using the product is consequential: if the AV does not properly perform its job, then it poses a mortal risk to the passenger(s). While there are several demographic variables that have been linked to willingness to adopt AVs—including youth, level of education, and tech savviness—trust appears to be the underlying psychological construct through which all of these variables ultimately impact willingness to adopt AVs (Haboucha et al., 2017; Lavieri et al., 2017; Menon et al., 2020).

In this work, however, we hypothesize that trust is not the main mechanism underlying patterns of liability ascription in the event of accidents where an AV is not at fault. Rather, we focus on the perceived relevance of counterfactual thinking. With that said, we do operationalize trust as an individual difference variable that may impact this mechanism. Specifically, we focus on individual differences in trust in an AV's driving competence, as opposed to other aspects of trust like integrity or values (Xie & Peng, 2009). We do this because consumers who share their opinions of AVs tend to raise negatives around malfunctions, fear, and loss of control, with 60% of one survey's respondents feeling "very concerned" about "computer system malfunctions causing a crash" (Bloomberg, 2016).

## The relevance of counterfactual thinking

Next, our proposed mechanism draws on work on counterfactual simulation. Consumer judgments are sometimes affected by counterfactuals (Folkes & Lassar, 2015; Tsiros & Mittal, 2000; Wiggin & Yalch, 2015)—psychological simulations of how events could have turned out differently, had an alternative course of action been taken (Kahneman & Tversky, 1982). To illustrate, participants in one seminal study read about a protagonist who used to take the same route to work every day, but 1 day decided to take a different, more scenic route instead (the "route" condition) before tragically being hit and killed by another driver who skipped a traffic light. When the authors asked participants to explain how things could have turned out differently, most cited the change in the protagonist's daily route, despite the many other causal explanations available. In short, participants tended to think of counterfactuals in which there was no deviation from what normally happens (Kahneman & Tversky, 1982).

More broadly, consumers tend to think of counterfactuals when an event is "abnormal," deviating from the statistical or social norm (Hitchcock & Knobe, 2009; Miller & McFarland, 1986; Phillips et al., 2015), and when factors of the event can easily be "mentally undone" as in "near miss" scenarios where the more favorable alternative seems to be in close proximity (Miller et al., 1990; Wiggin & Yalch, 2015). There are also individual differences in the propensity to think of counterfactuals (Kasimatis & Wells, 2014).

Within consumer psychology, counterfactual simulation has been implicated in a few notable domains, including: post-purchase regret and consumption choices (Roese et al., 2007; Strahilevitz et al., 2011; Tsiros & Mittal, 2000); responses to product breakdown and brand transgressions (Folkes & Lassar, 2015; Wiggin & Yalch, 2015); and promotion tactics and consumer responses to them (Krishnamurthy & Sivaraman, 2002; Li et al., 2022). In the current work, we explore whether and how the perceived relevance of counterfactual thinking affects product liability for a new technological product that is not yet a normal feature of most roads—autonomous vehicles, which involve surrendering control in a high-stakes context to artificial intelligence algorithms.

## Optimality bias in moral judgment

Finally, our proposed mechanism also draws on prior work showing an optimality bias in moral judgments (De Freitas & Johnson, 2018). When informed of a

harm that occurred, people do not merely entertain a counterfactual of what would "normally" happen on average, but they imagine what "ought to" or "should have" happened in the optimal scenario (De Freitas & Johnson, 2018; Phillips & Cushman, 2017). Notably, they expect agents to behave optimally even when it is unfair to have this expectation, as when blaming seismologists for failing to predict an earthquake (the optimal outcome) even when informed that the earthquake was impossible to predict (De Freitas & Johnson, 2018). The current work similarly tests whether, when participants perceive as relevant counterfactuals in which the driver of the not-at-fault vehicle was a human rather than an AV, they specifically consider as relevant the optimal scenario in which the driver avoided the accident—even when the scenarios are constructed so that a reasonable human driver could not have avoided the accident, as affirmed by experts (Studies 1 and 3) and/or ensured by the design of the scenario (Study 2).

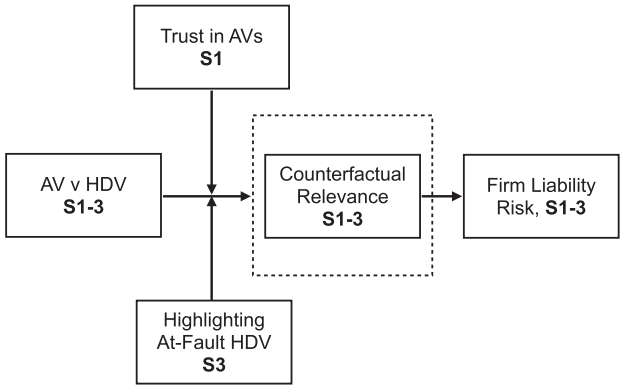


**FIGURE 1** Theoretical framework. Proposed thought process for assessing accidents not at fault, by which vehicle type influences whether consumers view as relevant a counterfactual in which the not-at-fault vehicle could have acted differently, which in turn informs judgments that the not-at-fault vehicle could have done more to avoid the accident and, ultimately, firm liability for the accident. Individual levels of trust in AVs influence counterfactual relevance. "S" refers to "study," indicating studies that test each component of the thought process.

## Theoretical framework

We expect that consumers view AVs as abnormal, unfamiliar, and unsafe and that these attitudes affect the counterfactuals that consumers believe are relevant when they learn about an accident involving an AV—even when the vehicle is not at fault.

After an accident occurs, we expect that consumers begin to search for a causal explanation (Kahneman & Tversky, 1982). If the not-at-fault vehicle is a human-driven vehicle (HDV), consumers primarily focus on the at-fault vehicle's responsibility for causing the accident, with little consideration of the not-at-fault HDV's role. However, when the not-at-fault vehicle is an autonomous vehicle (AV), the presence of the AV becomes salient due to its abnormality, given its occupant's lack of control. This abnormality prompts consumers to view as relevant a counterfactual scenario in which the vehicle could have acted differently. Because participants are inclined to view as relevant counterfactuals in which an agent acts differently, this leads them to assign liability to the AV manufacturer.

From a managerial perspective, we expect that liability ascriptions are additionally influenced by whether a company highlights the faults of the at-fault vehicle. Doing so may deflect attention away from the not-at-fault AV in the first place and back towards the at-fault vehicle, reducing consumers' tendency to view as relevant counterfactuals in which the AV could have acted differently. Thus, this attention-deflection intervention should reduce the perceived relevance of counterfactuals in the first place.

In short, this project extends existing theories of counterfactual thinking and optimality bias in moral judgment to new technology, and it proposes a model integrating both of these phenomena with AV adoption, wherein individual variation in attitudes towards AVs moderate the effects of counterfactual relevance on liability judgments Figure 1. This model provides a new theoretical lens for understanding the interplay between technology perception and legally relevant judgments.

## OVERVIEW OF STUDIES

We ask whether AV firms are viewed as more deserving of being sued than HDV firms for accidents not at fault, whether this response pattern is driven by the proposed counterfactual mechanism, and whether the effect of vehicle type on the relevance of counterfactual thinking is moderated by trust in AVs (Study 1). In addition to establishing the effect in a scenario modeled directly off of a real autonomous vehicle accident (Study 1), we replicate our effects using various realistic hypothetical scenarios (conceptual replications #1–4 in the Appendix S1). Next, providing more insight into the mechanism, we show that the counterfactual relevance mechanism is moderated by the particular counterfactuals that consumers spontaneously generate when prompted to do so (Study 2). Finally, we show that the effect is moderated by whether the at-fault vehicle's traffic-violating behavior is highlighted (Study 3). The university research ethics board approved the materials in all studies, and consent was obtained from all participants. Surveys, data, and code for all studies are included in the online GitHub repository for this project: https://github.com/Ethical-Intelligence-Lab/av_not_at_fault.

Altogether, we find robust evidence for all our conclusions from a total of 2677 participants. Together with our pre-study and four conceptual replications of Study 1,

we present evidence from 5834 participants. For generalization purposes, we sample participants from both the Mturk and Prolific platforms, and exclusion percentages never exceed 12%, apart from Study 1 which additionally excludes participants who recognized the real-world accident scenario upon which the study was based, for a total exclusion percentage of 20%.

The studies involve video and schematic recreations of accidents, inspired by the fact that such recreations are already at the heart of court cases involving AV-related accidents. Firms developing AV technology are using data recorders in their AVs in order to be able to reconstruct accident scenarios as a means of defending themselves in court and lowering insurance premiums, and in order to study and improve the driving skills of their AVs (AUVSI, 2012). Finally, all three of our driving scenarios have the same basic event structure, in which there is one vehicle at fault and one not at fault. The at-fault vehicle is always human-driven, while we vary whether the not-at-fault vehicle is an HDV or AV. For completeness, we compare liability ascriptions for not-at-fault AV manufacturers to all parties who could be held liable when the not-at-fault vehicle is human-driven, including the HDV manufacturer and not-at-fault human driver. However, since manufacturers and human drivers differ in several respects, for control sake, we only conduct mediation analyses for comparisons between AV vs. HDV manufacturers.

## STUDY 1

Study 1 tests whether the perceived liability of the manufacturer of a vehicle that is *not at fault* in an accident depends on whether it is human-driven or autonomous. In a pre-study, participants viewed AVs as less familiar, less safe, riskier, and more fear-inducing than HDVs, showing that AVs are perceived as more abnormal and threatening on several dimensions as compared to HDVs (see Appendix S1). Because AVs violate the norm in which a human is in control of the vehicle, we predict that when an accident occurs and the not-at-fault vehicle is an AV, the not-at-fault vehicle becomes more salient to consumers. Because of this, participants are more likely to view as relevant a counterfactual in which the vehicle had acted differently, avoiding the accident. Given this thought process, they infer that the firm is, therefore, partially liable for the damages.

We also test whether there is a greater willingness to view the manufacturer of an AV as liable as compared to a *human driver* of a HDV, who should be viewed as just as agentic as, if not more agentic than, the AV manufacturer.

Furthermore, we investigate whether individual differences in trust towards AVs—as measured via an existing psychological scale, modified for AVs (Moorman et al., 1992)—moderate these effects.

We test a scenario directly lifted from a news report of an accident involving an autonomous vehicle provided by Cruise, a subsidiary of General Motors. In 2023, a pedestrian was hit by a Nissan driven by a person who did not brake, and then thrown into the path of a Cruise vehicle. Despite the Cruise vehicle's attempt to brake (Lawyers, 2024), it collided with the pedestrian. An independent engineering firm determined that a human driver in the same situation would not have been able to avoid the crash (Cano, 2024). Despite the fact that Cruise was deemed not at fault, its driving license in San Francisco was permanently suspended by the Department of Motor Vehicles, in part because of how the vehicle behaved after the accident and how the company interacted with regulators and the media (Bensinger, 2024; Cano, 2024). Even so, a natural question raised by our conceptual model is whether Cruise was penalized more heavily merely because its vehicle was autonomous.

### Method

This study was pre-registered (https://aspredicted.org/3Q2_GYD). Due to potential concerns around the discriminant validity of the proposed serial mediators ($r$=0.73 in Study 1), we ran a simpler mediation model with counterfactual relevance as the sole mediator rather than our original plan to run a serial mediation. Except for this change, we did not deviate from the preregistered plan.

We recruited 895 participants (US residents only) from Amazon's Mechanical Turk, who passed attention checks and completed the survey, in exchange for $0.50. We excluded 176 for failing comprehension checks or for recognizing the scenario as the Cruise incident (described below), yielding 719 participants ($M_{age}$=42.0, 58.1% females). Participants were only allowed to participate if they correctly answered two attention checks at the beginning of the survey.

Participants were assigned to one of two conditions (agent: AV or HDV) in a between-subjects design. Participants first rated how much they trusted AVs. To this end, we utilized five statements from an existing psychological scale originally developed to measure trust between managers and researchers (Moorman et al., 1992), adapting it to refer to AVs. We found the original scale fitting for the AV context, because it assessed managers' beliefs in researchers' *competence*, while in this study we intended to measure trust in the technology's competence. In the current study, participants indicated the extent to which they agreed with the following statements on a scale anchored from 0 (Completely disagree) to 100 (Completely agree), presented in randomized order: (1) I would be willing to let an AV make important driving decisions without my involvement; (2) If I was unable to monitor my driving

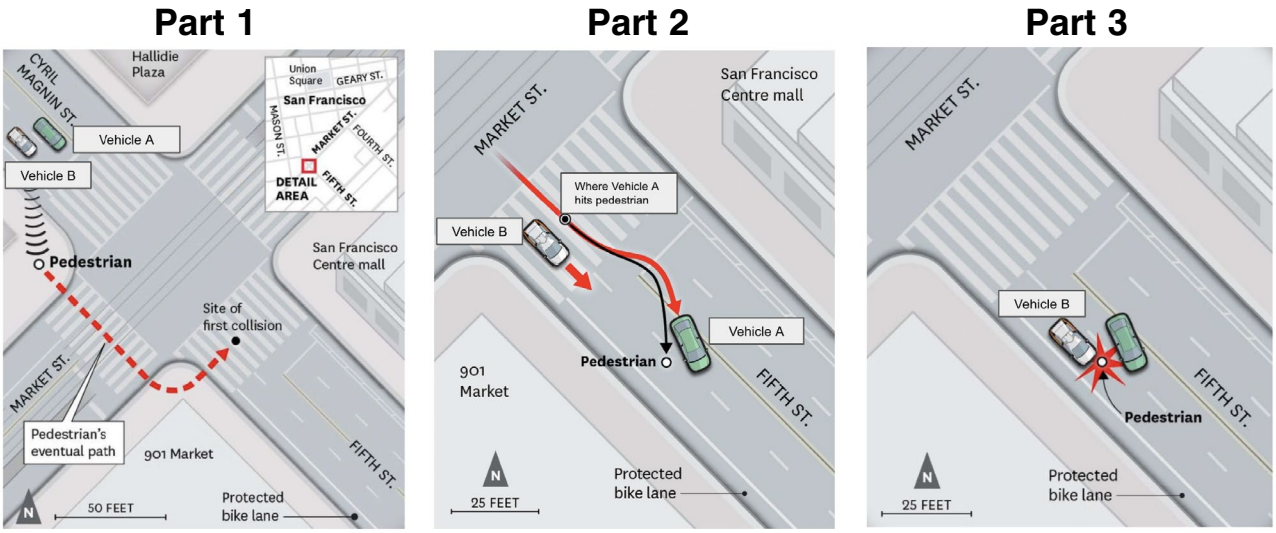


**FIGURE 2** Scenario and schematics in Study 1. Images and crash details taken from Cano (2024).

**TABLE 1** Measures in Study 1.

| Measure | Statement |
|---|---|
| Reasonable to sue, At Fault (DV) | *It would be reasonable to sue the driver of Vehicle A to cover the costs of the serious injuries sustained by the pedestrian* |
| Reasonable to sue, Not at Fault (DV) | *It would be reasonable to sue the manufacturer of Vehicle B to cover the costs of the serious injuries sustained by the pedestrian* |
| Reasonable to sue, Not at Fault[a] (DV) | *It would be reasonable to sue the manufacturer of Vehicle B to cover the costs of the serious injuries sustained by the pedestrian* |
| Counterfactual relevance, At Fault (M) | *When it comes to thinking about how the injury could have been avoided, it is relevant to consider what Vehicle A could have done differently* |
| Counterfactual relevance, Not at Fault (M) | *When it comes to thinking about how the injury could have been avoided, it is relevant to consider what Vehicle B could have done differently* |
| Done more, Not at Fault (M) | *Vehicle B could have done more to avoid the accident* |

Abbreviations: DV, dependent variable; M, mediator; MOD, moderator.
[a]Only measured in HDV condition.

activities, I would be willing to trust an AV to get the job done right; (3) I trust an AV to do things I can't do myself; (4) I trust an AV to do things my vehicle can't do itself; (5) I generally do not trust an AV.

Next, participants were told that they would read an excerpt from a news article describing a crash involving the vehicle in question. Both the excerpt and schematics were taken from an article in the San Francisco Chronicle (Cano, 2024)—Figure 2.

Participants then indicated the extent to which they agreed with several statements anchored on scales from 0 (Completely disagree) to 100 (Completely agree) and presented in randomized order. Each statement was presented on its own page, accompanied by the schematics in Figure 2. The dependent variables pertained to whether participants thought it would be reasonable to sue the at-fault driver and manufacturer of the not-at-fault vehicles, although we were chiefly interested in the latter (Table 1). First, the DV measures were presented in randomized order. Next, the counterfactual relevance mediator was presented in randomized order. Finally, the measure of Vehicle B's ability to do more was presented.

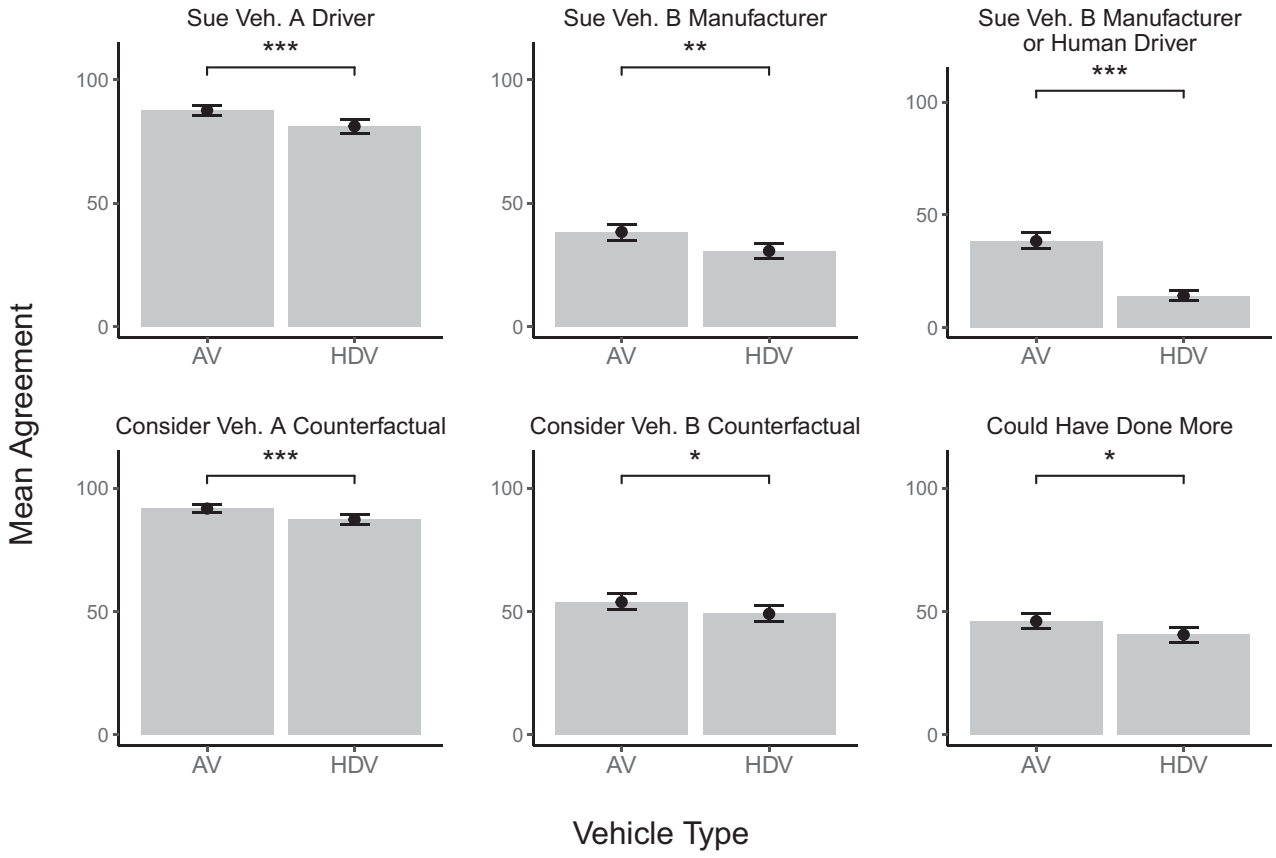


**FIGURE 3** Main results for Study 1. *$p < 0.05$. **$p < .01$. ***$p < .001$.

**TABLE 2** Statistics for Study 1.

| Measure | AV mean | HDV mean | *T*-value | Cohen's *D* |
|---|---|---|---|---|
| Sue, At Fault (DV) | 87.60 (19.93) | 81.14 (26.48) | $t(661) = 3.69$*** | 0.28 |
| Firm Sue, Not at Fault (DV) | 38.33 (32.57) | 30.63 (31.09) | $t(716) = 3.24$** | 0.24 |
| Counterfactual, At Fault (M) | 91.73 (14.81) | 87.29 (19.49) | $t(664) = 3.43$*** | 0.26 |
| Counterfactual, Not at Fault (M) | 53.83 (32.10) | 49.06 (31.60) | $t(717) = 2.01$* | 0.15 |
| Done more, Not at Fault (M) | 46.08 (31.78) | 40.61 (30.55) | $t(717) = 2.36$* | 0.18 |
| Trust in Autonomous Vehicles | 34.35 (23.94) | 31.81 (25.25) | $t(714) = 1.38$ | 0.10 |

*Note*: *T*-statistics reflect results of independent-samples *t*-tests.
*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$.

After completing the measures, participants answered two comprehension checks about what type of vehicle they saw in the scenario (AV or HDV) and which vehicle hit the pedestrian first (vehicle A or B) and responded to a Yes/No question ("When you read the scenario, did you recognize that it was describing a real incident involving the company Cruise?") as a measure of whether they recognized the scenario as the Cruise incident. Finally, participants answered standard demographics questions. Participants who failed either of the comprehension checks or who recognized the incident were excluded from the analysis.

## Results

For each of the measures that were completed in both conditions, we compared the measure between AVs and HDVs not at fault, finding significant differences for all measures except trust (Figure 3, Table 2).

Firstly, participants thought it was very reasonable to sue the human driver of the at-fault vehicle — in this case, the first vehicle that struck the pedestrian — in both the AV and HDV conditions. Notably, participants also thought it was more reasonable to sue the manufacturer of a not-at-fault vehicle when the

not-at-fault vehicle was an AV than either the manufacturer ($M_{AV}=38.33$ $M_{HDV}=30.63$, $t(716)=3.24$, $p=0.001$, $d=0.24$) or driver ($M_{AV}=38.33$ $M_{HDV}=14.08$, $t(642)=11.63$, $p<0.001$, $d=0.87$) of the same vehicle when it was a conventional HDV. These results suggest that manufacturers of AVs face a higher liability risk, even in accidents where they are not at fault and where manufacturers of HDVs and human drivers of HDVs would be judged more favorably. In line with the hypothesized thought process, we found that participants were more likely to view the counterfactual as relevant (in which Vehicle B had behaved differently) when Vehicle B was an AV than an HDV (Table 2).

The AV trust measure was averaged across all five trust measures ($\alpha=0.92$). There was no difference in trust between conditions (Table 2).

## Mediation analysis

In the Appendix S1, we report correlation tables between all variables, showing an indication of discriminant validity between our DV and intended mediator ($r=0.56$). We conducted a mediation analysis to determine whether the hypothesized causal order: condition → counterfactual → reasonable to sue, explains our findings. The mediation was statistically significant for the manufacturer comparison ($b=-2.65$, SE = 1.34, 95% CI [−5.26, −0.07]).

Next, we conducted a moderated mediation analysis (PROCESS Model 7; Hayes, 2012). This model featured the same mediation described above, with reasonable to sue as the DV, but with the "A" path between vehicle condition and counterfactual relevance moderated by baseline trust in AVs. The index of moderated mediation was significant ($b=0.25$, SE = 0.06, 95% CI [0.13, 0.37]). The less participants trusted AVs at baseline, the more they considered the counterfactual as relevant (Figure 4). The full mediation diagram for this and subsequent experiments is provided in the Appendix S1.

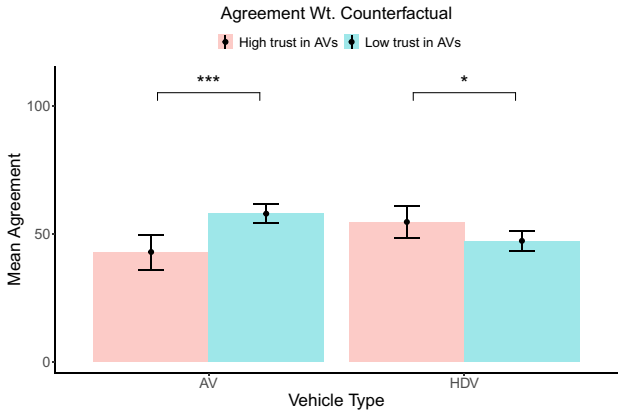


**FIGURE 4** The effect of trust in AVs on agreement with the counterfactual relevance mediator, showing the moderation pattern in Study 1. High Trust >50 average on the trust scale; Low Trust ≤ 50. Error bars indicate 95% CIs. *$p<0.05$, ***$p<0.001$.

We additionally tested an alternative hypothesis wherein the effect of vehicle type on manufacturer liability is instead mediated through trust in AVs. We conducted a mediation analysis (PROCESS Model 4; Hayes, 2012) testing for the following path: condition → trust in AVs → DV. We found this alternative mechanism was not statistically significant ($b=0.40$, SE = 0.33, 95% CI [−0.15, 1.15]), indicating that patterns of liability are not explained by trust as a mediating mechanism.

For all three studies, we also explored whether the effect of vehicle type on manufacturer liability is moderated by participant age, given the potential for older people to be more resistant to AV technology (Park & Han, 2023). Using simple linear regression, we did not find statistically significant moderation effects for age across the three studies.

## Discussion

Using a scenario directly lifted from news coverage involving a real autonomous vehicle, we find that consumers view AV manufacturers as more liable than HDV manufacturers for the same accident, rooted in the proposed "relevance of counterfactual thinking" mechanism. The findings suggest that this mechanism may come into play whenever real autonomous vehicles are involved in accidents and that it may have already influenced the heavy penalties that Cruise faced after the accident. That is, regardless of the explicit reasons offered for removing Cruise's license to operate, these punishments may have been driven in part by the underlying intuition that the vehicle was defective because a human in the same position could have acted differently, avoiding the accident. Additionally, we found that the degree to which consumers viewed counterfactuals as more relevant for AVs than HDVs was moderated by their trust in AVs. However, trust alone did not explain the main differences found in the study.

In the Appendix S1 we report four conceptual replications of these results. The first replication study shows similar results with a different outcome measure of the perceived liability of the firm. The second replication implicates the same counterfactual mechanism in judgments that the vehicle is defective, which is the legal requirement for holding a not-at-fault party liable. The third replicates the results in a third driving scenario. The fourth replicates the results in a fourth driving scenario while testing whether individual differences in political affiliation and perceived driving ability moderate these effects. We expected that conservatives, being more averse to new AI technology (Castelo & Ward, 2021), would be more inclined to view counterfactuals as relevant for AVs, and we expected that those who think they

are better drivers than average (Shariff et al., 2021) would be more likely to hold firms as liable. In both cases, we found that the index of moderated mediation for the moderators was not significant.

## STUDY 2

In order to gain more insight into the validity of the counterfactual relevance mechanism, Study 2 prompts participants to generate counterfactuals and investigates whether specifically generating counterfactuals around the not-at-fault vehicle predicts patterns of liability ascriptions to the firm. We anticipate that (i) participants will be more likely to generate counterfactuals that suggest the not-at-fault vehicle could have acted differently when the vehicle is an AV than an HDV, and (ii) this tendency will predict higher firm liability when the vehicle is an AV. In addition, we explore the nature of the counterfactuals generated around the not-at-fault vehicle in the AV condition. Finally, this study gathers further information to assess discriminant validity by measuring the mediator and outcome constructs with two items each.

### Method

Informed by a pilot study using a similar method to the current study, we recruited 422 participants from Amazon's Mechanical Turk, who passed attention checks and completed the survey, in exchange for $0.50. We excluded 65 for failing the same checks as Study 1 as well as one participant for whom we did not reach inter-coder reliability (described below), yielding 357 participants ($M_{age}$=43.2, 54.3% females).

Participants were assigned to one of two conditions (agent: AV or HDV) in a between-subjects design. The scenario and checks were identical to Study 1. This time, before answering questions, all participants were given the counterfactual prompt from Kahneman and Tversky (1982), minimally adapted for the current scenario: "As commonly happens in such situations, the family of the pedestrian often thought and often said 'If only…', during the days that followed the accident. How did they continue this thought? Please write two or more likely completions."

Participants answered the same measures as Study 1, except we measured the counterfactual relevance construct with two items ("How much do you agree: It's important to consider whether Vehicle B could have acted differently when thinking about how the injury could have been prevented." and "How much do you agree: When it comes to thinking about how the injury could have been avoided, it is relevant to consider what Vehicle B could have done differently."), as we did the liability risk construct ("How much do you agree: It would be reasonable to sue the driver of Vehicle B to cover the costs of the serious injuries sustained by the pedestrian." and "How much do you agree: The manufacturer of vehicle B is liable for the serious injuries sustained by the pedestrian."). Given these additional items, we also removed two items that were not needed (about whether Vehicle B could have avoided the accident, and whether it was reasonable to sue the human driver of the not-at-fault vehicle in the HDV condition).

### Results

We averaged the items measuring counterfactual relevance ($r$=0.83) and liability risk ($r$=0.86) into single constructs, given high reliability. We assessed discriminant validity between our counterfactual relevance and liability risk items using the Heterotrait-Monotrait (HTMT) ratio, using the recommended threshold of 0.85 (Henseler et al., 2015). The HTMT value was 0.48, supporting the discriminant validity of the constructs.

Next, two independent, condition-blind coders (Julian De Freitas and Xilin Zhou) coded the counterfactual explanations based on whether they suggested blame of the not-at-fault vehicle (see Appendix S1 for full instructions), following the exact procedure recommended by Rhee et al. (1995); inter-coder reliability was high ($r$=0.99). As predicted, participants were more likely to mention the not-at-fault vehicle when in the AV than HDV condition (43.18% versus 14.36%, $\chi^2(1, N=357)=34.91$, $p<0.001$). For the remainder of the analysis, we treat these codes as a quasi-experimental variable.

We ran a 2 (agent: AV or HDV) ×2 (mentions not-at-fault vehicle: yes or no) ANOVA for the liability risk DV. We found main effects of agent type ($M_{AV}$=31.63, $M_{HDV}$=13.01, $F(1, 353)=48.81$, $p<0.001$, $\eta^2=0.12$) and mentioning the not-at-fault vehicle ($M_{yes}$=38.55, $M_{no}$=15.64, $F(1, 353)=34.20$, $p<0.001$, $\eta^2=0.09$), and an interaction effect ($F(1, 353)=29.04$, $p<0.001$, $\eta^2=0.08$). Participants thought it was more reasonable to sue the manufacturer when the not-at-fault vehicle was an AV than an HDV, but only when their counterfactual mentioned the not-at-fault vehicle ($M_{AV}$=48.80, $M_{HDV}$=8.60, $t(99)=8.31$, $p<0.001$, $d$=1.28) as opposed to when it did not ($M_{AV}$=18.57, $M_{HDV}$=13.75, $t(184)=1.62$, $p$=0.11, $d$=0.22). Thus, the more that participants thought of counterfactuals involving the not-at-fault AV (but not HDV), the more they viewed its manufacturer as liable. Table 3 summarizes all pairwise comparisons.

We additionally ran an ANOVA for our measure of counterfactual relevance. Unlike Study 1 and the four conceptual replications in the Appendix S1, we find no main effect of agent type ($M_{AV}$=49.05, $M_{HDV}$=49.00, $F(1, 353)=0.00$, $p$=0.99, $\eta^2=0.00$), although we find a main effect of mentioning the not-at-fault vehicle ($M_{yes}$=62.06, $M_{no}$=43.81, $F(1, 353)=32.14$, $p<0.001$, $\eta^2=0.08$), and an interaction effect ($F(1, 353)=10.72$, $p$=0.001, $\eta^2=0.03$).

**TABLE 3** Statistics for Study 2.

| Measure | AV v HDV not-At-fault mentioned | AV v HDV not-At-fault not mentioned |
|---|---|---|
| Sue, At Fault (DV) | $t(46) = 0.39$, $d = 0.09$ | $t(241) = 2.55$*, $d = 0.31$ |
| Firm Liability, Not at Fault (DV) | $t(99) = 8.31$***, $d = 1.28$ | $t(184) = 1.62$, $d = 0.22$ |
| Counterfactual, At Fault (M) | $t(39) = 1.04$, $d = 0.25$ | $t(233) = 2.27$*, $d = 0.28$ |
| Counterfactual, Not at Fault (M) | $t(41) = 1.96$, $d = 0.46$ | $t(216) = -3.16$**, $d = -0.40$ |

*Note*: *T*-statistics reflect results of independent-samples *t*-tests.
*$p<0.05$. **$p<0.01$. ***$p<0.001$.

Whether participants mentioned the not-at-fault vehicle did not affect ratings of counterfactual relevance in the HDV condition ($M_{yes} = 52.40$, $M_{no} = 48.43$, $t(34) = -0.63$, $p = 0.53$, $d = 2.29$), perhaps because participants thought there was nothing abnormal about that vehicle to begin with. In contrast, mentioning the not-at-fault vehicle did affect ratings of counterfactual relevance in the AV condition, with those mentioning the not-at-fault vehicle finding it very relevant to consider how it could have acted differently and those who did not feeling otherwise ($M_{yes} = 65.37$, $M_{no} = 36.65$, $t(164) = -6.70$, $p < 0.001$, $d = 2.13$); thus, there was a "polarizing" effect relative to the overall HDV mean ($M = 49.00$), explaining the null effect of vehicle type on counterfactual relevance. Perhaps in this AV condition, it was already the default to consider a counterfactual involving the not-at-fault vehicle, such that explicitly mentioning it was a good indicator of this tendency (and vice versa for not explicitly mentioning it).

### Exploratory analysis

To better understand the nature of the counterfactuals generated around the not-at-fault AV, we also coded the types of explanations provided in the AV condition when participants mentioned the not-at-fault vehicle into three categories (see Appendix S1 for full coding instructions): Saying that things would have turned out differently if (i) the vehicle had behaved or were designed differently, (ii) there had been a human in the car, and (iii) the vehicle did not exist or AVs were not permitted in the first place.

In this condition, 56.63% mentioned a human being in the car instead (e.g., "If only there were a person in the car they would have been able to swerve avoiding my family member" or "If only a human more aware than a computer had been driving the second vehicle."). 30.26% mentioned the vehicle acting or being designed differently (e.g., "If only the car had been programmed to prioritize the safety of pedestrians over the convenience of the vehicle's passengers" or "If only Vehicle B had a faster pedestrian response road system."). 17.11% mentioned the vehicle not being there at all or AVs being disallowed (e.g., "If only self-driving cars didn't exist" or "If only the autonomous vehicles were prohibited"). These proportions differed from chance ($\chi^2(2, N = 357) = 14.71$, $p < 0.001$). There was no difference in liability risk for the AV manufacturer depending on which counterfactual explanations about the not-at-fault AV were provided ($F(1, 73) = 0.75$, $p = 0.48$, $\eta^2 = 0.02$).

## Discussion

We found that counterfactual explanations that mentioned the not-at-fault vehicle were associated with higher liability for AVs than HDVs. This result strengthens evidence for the proposed counterfactual relevance mechanism while showing that the actual content of the counterfactuals provides an explanatory role.

At the same time, while the main effects strongly suggest a causal role of counterfactual content on liability perceptions, we note that strong claims about causality are tempered by the fact that counterfactual explanations were (by necessity of our design) a quasi-experimental condition generated by participants rather than experimentally manipulated. This was a deliberate design choice, as asking people to think of certain counterfactuals could contaminate the results, that is, we would not know whether those would be the counterfactuals that they would otherwise spontaneously generate.

Tellingly, our exploratory analysis found that consumers are particularly inclined to imagine counterfactuals in which a human was in the vehicle instead, speaking to the idea that they were most focused on what is abnormal and striking about these vehicles: the lack of a human who is in control of the vehicle.

# STUDY 3

Inspired by the results of Study 2, Study 3 investigates whether we can intervene on the main effect by deflecting the participant's attention away from the not-at-fault vehicle in the first place, to focus it instead on the faults of the *at-fault* vehicle. Our theoretical framework posits that whereas participants in the HDV condition primarily focus on the at-fault vehicle's responsibility, participants in the AV condition instead turn their attention toward the abnormality of the AV's presence. Thus, deflecting attention to the faults of the at-fault human driver should make it less likely that participants will focus on this abnormal presence of the AV, making the counterfactual seem less relevant in the first place.

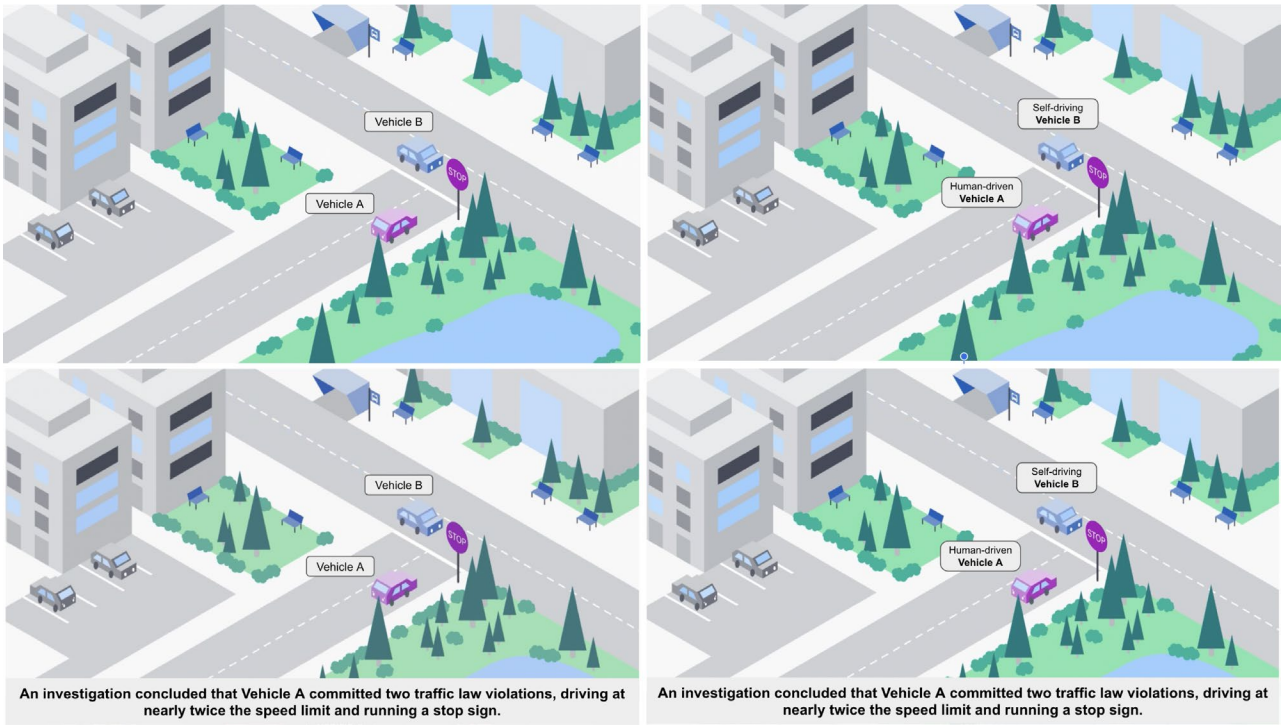


**FIGURE 5** Instruction stills for agent (human-driven, left column; autonomous, right column) and intervention (no intervention, top row; intervention, bottom row) conditions.

In short, we predicted that the deflection manipulation would moderate the first path of the conceptual model, as trust did in Study 1 and mentions of the not-at-fault vehicle did in Study 2. To ensure further robustness, we also use another driving scenario, although we used the same measures as Study 1.

## Methods

This study was pre-registered (https://aspredicted.org/NJ1_X1D), and we did not deviate from the pre-registered plan except that we again ran a simpler mediation with counterfactual relevance as the sole mediator rather than our original plan to run a serial mediation, given discriminant validity concerns about the serial mediators ($r = 0.63$). We recruited 1360 participants from Amazon's Mechanical Turk, who passed attention checks and completed the survey, in exchange for $0.70. We excluded 121 for failing similar comprehension checks as Studies 2 (described below), yielding 1239 participants ($M_{age} = 43.7$, 50.5% females).

Participants were assigned to one of four conditions in a 2 (agent: AV or HDV) x 2 (intervention: yes or no) between-subjects design. Participants in all conditions were given the following instructions (information in squared parentheses was only included in the AV condition) accompanied by Figure 5:

> On the next page, you will watch an animated video of a traffic scenario, depicted below. The video shows an intersection in which the driver of Vehicle A runs a stop sign and strikes Vehicle B, seriously injuring its occupant. [Vehicle B is a fully autonomous self-driving car, which means that it is driven by a computer algorithm, and its human occupant has no control of the vehicle.]

In the intervention condition, the instructions additionally included the text: "An independent accident investigation concluded that Vehicle A committed two traffic law violations, driving at nearly twice the speed limit and running a stop sign." In all conditions, the instructions were accompanied by the relevant image from Figure 5.

Readers can view the video here, which was the same across conditions: http://y2u.be/lPRH7NTtF8M. Participants watched the video twice, after which those in the intervention condition were asked the following comprehension check to ensure their understanding of the traffic violations: From the options below, please select the two traffic violations which Vehicle A committed: [Driving at nearly twice the speed limit; Running a stop sign; Passing a school bus]. After completing the same main measures as in Study 1, participants were asked to answer two comprehension checks about what

type of vehicle they saw in the scenario (AV or HDV) and which vehicle ran a stop sign (vehicle A or B).

## Results

We ran 2 (agent: AV or HDV) × 2 (intervention: yes or no) ANOVAs for each of our two primary DVs (manufacturer liability; AV manufacturer vs. human driver liability). For manufacturer vs. manufacturer liability, we found a main effect of agent type ($M_{AV}$ = 14.59, $M_{HDV}$ = 10.40, $F(1, 1235) = 11.89$, $p < 0.001$, $\eta^2 = 0.010$), no main effect of intervention ($M_{intervention}$ = 11.82, $M_{no\ intervention}$ = 13.26, $F(1, 1235) = 1.42$, $p = 0.23$, $\eta^2 = 0.001$), and an interaction effect ($F(1, 1235) = 7.94$, $p = 0.005$, $\eta^2 = 0.006$). As in prior studies, participants were more likely to think it was reasonable to sue the AV than HDV manufacturer when there was no intervention ($M_{AV}$ = 17.00, $M_{HDV}$ = 9.38, $t(554) = 4.43$, $p < 0.001$, $d = 0.35$). However, when participants received additional details about the at-fault vehicle's traffic violations, there was no longer a significant difference in these ratings between the AV and HDV conditions ($M_{AV}$ = 12.20, $M_{HDV}$ = 11.43, $t(617) = 0.45$, $p = 0.65$, $d = 0.04$). The intervention successfully eliminates the main effect found in Study 1. Table 4 summarizes all pairwise comparisons.

For AV manufacturer vs. human driver reasonableness to sue, we found a main effect of agent type ($M_{AV}$ = 14.59, $M_{HDV}$ = 6.36, $F(1, 1235) = 43.16$, $p < 0.001$, $\eta^2 = 0.034$), no main effect of intervention ($M_{intervention}$ = 9.60, $M_{no\ intervention}$ = 11.52, $F(1, 1235) = 2.41$, $p = 0.12$, $\eta^2 = 0.002$), and an interaction effect ($F(1, 1235) = 5.42$, $p = 0.020$, $\eta^2 = 0.004$). Participants were more likely to judge it was reasonable to sue the AV manufacturer than the human driver, both without the intervention ($M_{AV}$ = 17.00, $M_{HDV}$ = 5.85, $t(591) = 6.18$, $p < 0.001$, $d = 0.49$) and with the intervention ($M_{AV}$ = 12.20, $M_{HDV}$ = 6.88, $t(615) = 3.08$, $p = 0.002$, $d = 0.25$), although the effect was smaller with the intervention.

**TABLE 4** Statistics for Study 3.

| Measure | AV v HDV No intervention | AV v HDV with intervention |
|---|---|---|
| Sue, At Fault (DV) | $t(569) = -2.36$*, $d = -0.19$ | $t(617) = -0.24$, $d = -0.02$ |
| Sue Firm, Not at Fault (DV) | $t(554) = 4.43$***, $d = 0.35$ | $t(617) = 0.45$, $d = 0.04$ |
| Counterfactual, At Fault (M) | $t(617) = -0.65$, $d = -0.05$ | $t(548) = 0.63$, $d = 0.05$ |
| Counterfactual, Not at Fault (M) | $t(578) = 6.16$***, $d = 0.49$ | $t(618) = 2.51$*, $d = 0.20$ |
| Done more, Not at Fault (M) | $t(600) = 5.71$***, $d = 0.46$ | $t(588) = 4.29$***, $d = 0.34$ |

*Note*: *T*-statistics reflect results of independent-samples *t*-tests.
*$p < 0.05$. ***$p < 0.001$.

We additionally ran an ANOVA for our measure of counterfactual relevance, finding a main effect of agent type ($M_{AV}$ = 29.23, $M_{HDV}$ = 19.29, $F(1, 1235) = 36.11$, $p < 0.001$, $\eta^2 = 0.028$), no main effect of intervention ($M_{intervention}$ = 24.10, $M_{no\ intervention}$ = 24.63, $F(1, 1235) = 0.11$, $p = 0.74$, $\eta^2 = 0.000$), and an interaction effect ($F(1, 1235) = 5.52$, $p = 0.019$, $\eta^2 = 0.004$). Participants were more likely to rate the counterfactual as relevant when the not-at-fault vehicle was an AV, both without the intervention ($M_{AV}$ = 31.42, $M_{HDV}$ = 17.59, $t(578) = 6.16$, $p < 0.001$, $d = 0.49$) and with the intervention ($M_{AV}$ = 27.05, $M_{HDV}$ = 21.00, $t(618) = 2.51$, $p = 0.012$, $d = 0.20$), although the effect was smaller with the intervention.

### Mediation results

In the Appendix S1, we report correlation tables between all variables, finding an indication of discriminant validity between our DV and intended mediator ($r = 0.29$).

As pre-registered, we tested whether the intervention condition moderates the "A" path of the model between vehicle type condition and counterfactual relevance (PROCESS Model 7; Hayes, 2012). The index of moderated mediation was found to be significant for the manufacturer vs. manufacturer comparison, ($b = -1.60$, SE = 0.73, 95% CI [−3.11, −0.26]). This indicates that the tested intervention successfully reduces the relevance of counterfactual thinking around what the AV could have done differently, in turn decreasing participants' tendencies to find AV manufacturers more liable.

## Discussion

We found that emphasizing the fault of the *at-fault* vehicle by providing additional details on its traffic violations significantly attenuates the liability risk for AV firms (vs. HDV firms and human drivers). Consistent with our account that such an intervention deflects attention away from the AV, the intervention reduced the perceived relevance of the counterfactuals in the first place. This result both strengthens the evidence for our theoretical framework and suggests a managerially actionable intervention. In the wake of accidents where their vehicles are not at fault, AV manufacturers may want to focus their communications on the fault of the at-fault driver. Likewise, they may want to do the same in court, if the accident leads to a trial.

# GENERAL DISCUSSION

Across three studies and four supplemental studies, we found that vehicle manufacturers are more likely to incur liability risk when their vehicles are autonomous than when they are human-driven, in the event that

their vehicles are not at fault. This response pattern was driven by a tendency to consider as relevant counterfactual scenarios in which the AV could have acted differently, and conclude that the accident could have been avoided; hence, consumers thought it was more reasonable to sue or view as liable an AV firm than an HDV firm in the same scenario. Similarly, an AV firm incurred more liability risk than a human driver of a not-at-fault HDV, showing that in practice it incurred greater liability risk than all comparable parties. Liability-related ascriptions were not explained by levels of trust in AVs, although the less participants trusted AVs, the more relevant they thought it was to consider as relevant counterfactuals in which the AV acted differently. Likewise, liability-related ascriptions were not merely explained by the expectation that AVs should drive better than humans.

## Theoretical implications

Our research has three main theoretical implications. First, we contribute to work on consumer reactions to AI failures, which typically find lower blame for AI than humans, driven by inferences of lower agency for AI (Arikan et al., 2023; Li et al., 2016; Srinivasan & Sarial-Abi, 2021). We find the reverse effect in scenarios where AI is not at fault, due to a distinct counterfactual relevance mechanism in which "abnormal" AVs trigger the perceived relevance of counterfactuals in which they act more optimally, leading to *higher* liability for AV firms. We expect that the same process may be at play for other new AI technologies that are viewed as abnormal, such as new chatbots. Furthermore, while previous work on AI failures in the context of AVs has focused on hypothetical moral dilemmas in which AVs are forced to make a choice between two harmful outcomes (Awad et al., 2018; Bonnefon et al., 2016; Gill, 2020), the current work suggests economic and social risks arising from how consumers think about the exceedingly more prevalent scenario in which AVs are involved in accidents not at fault.

Second, we contribute to work on the role of trust in the adoption of AVs (Gold et al., 2015; Xu et al., 2018). Although trust is directly relevant to AV adoption, we find that in the case of AV liability for accidents not at fault, trust is mostly indirectly relevant by moderating the extent to which consumers view as relevant counterfactuals in which a human would act more optimally.

Third, we contribute to work on counterfactual reasoning in consumer psychology, by revealing that new technology affects which counterfactuals are viewed as relevant in event-based scenarios, and that this influences inferences about product defects and firm liability.

## Practical implications

If AV firms incur greater liability risk than HDV firms for identical accidents not at fault, then this outcome has both economic and societal implications. Economically, the lawsuits from these accidents may be prohibitively expensive for AV firms and their investors, given the costs of settlements, claims administration costs, and legal fees for each claim filed (Morgan, 1982). This suggests that even though the size of the overall "pie" of accidents is expected to be smaller for AVs, firms may be responsible for a larger "slice" of that pie than they are currently (Smith, 2017). Societally, if firms must charge higher prices to cover higher anticipated liability costs, as via higher ride-sharing prices, this may discourage adoption and ultimately delay the progress of this technology and reduce the expected prevention of accident-related injuries and deaths (Nichols, 2013; Villasenor, 2014). In the extreme, firms and investors may avoid AVs altogether.

In fact, the risks we have uncovered here might be magnified, for several reasons. First, when bringing Products Liability, General Liability, or Auto Liability claims against a defendant in some states of the United States, the liability amount is un-capped and, as a form of punishment, can exceed the estimated cost of damages (Moulton, 2019). If the public places undue liability on AVs, then Plaintiffs could appeal to their misguided perceptions to seek higher rewards. Second, any unanticipated public perception or litigation risks stemming from not-at-fault accidents are heightened by the higher frequency of such accidents (as compared to accidents where the AV is at fault), with potentially large and unexpected financial impacts on the bearers of risk (e.g., insurers or a self-insured company). Third, because AV firms are more likely to have the means to cover damages than individual human parties, this may make them attractive targets of lawsuits, even if they are viewed as just weakly or partially liable (Smith, 2017). Relatedly, under the law of "joint and several liability" that is active in some states of the US, a party that is only partially responsible for the damages may be required to pay all damages if they are the only party carrying insurance (Wright, 1992). Fourth, because some bearers of risk will increase insurance premiums to account for the liability risks, some firms may choose to not take on insurance at all, exposing themselves to potentially extraordinary risk if a costly lawsuit is brought against them.

To proactively avert or at least minimize these risks, all stakeholders may want to normalize AVs in the minds of consumers. A possible silver lining is that once AVs are rolled out and become ubiquitous in large cities, the feeling that they are abnormal should be reduced. At the same time, it is yet unclear what kinds of exposure to AVs will have this effect, and how long it will take to reach an equilibrium in which AVs seem as normal as HDVs.

Similarly, deployments could be delayed if consumers are exposed to news that confirms their current mistrust—as in various recent reports that the technology has been over-hyped (Chafkin, 2022; Isidore et al., 2022), public campaigns against the technology (Vynck, 2022), and salient accidents involving AVs—even if accidents are rarer for AVs than HDVs. This dynamic will continue to play out in the early days of adoption, with potential long-term consequences for whether the technology is widely adopted.

Finally, AV manufacturers can also reduce the perceived liability of the not-at-fault AV by highlighting the faults of the at-fault driver, as through news communication efforts or in its defense (if the accident goes to trial). Such an intervention works by reducing the perceived relevance of counterfactuals involving the not-at-fault AV, presumably because it deflects attention and directs blame to the at-fault party.

## Limitations and future directions

Our findings raise several open questions for future work on psychological mechanisms, generalization of the effects, and potential interventions.

On psychological mechanisms, one question is whether the same thought process at play here affects not just views about firm liability but also brand image, with effects on purchase intent and word of mouth. Future investigations can also probe whether other psychological processes contribute to the liability response patterns found here. One possibility is that consumers employ a generalization heuristic, assuming that an error with one AV also implicates all other AVs from the firm or even all AVs in general, resulting in a larger total risk of harm (as in so-called "algorithmic transference" effects; Longoni et al., 2022). If such an inference is at play, it would be in addition to the counterfactual relevance mechanism uncovered here, which was causally implicated in the response pattern. Future work could also further probe whether negative attitudes towards AVs, such as believing that they will eliminate human jobs, affect liability ascriptions in addition to the counterfactual relevance process uncovered here.

On generalization, future studies can expand upon the liability and insurance risks for firms by surveying other relevant stakeholders, such as underwriters and lawyers. It can also measure how consumers apportion liability across various stakeholders in the value chain, such as vehicle manufacturers, software providers, and bus operators. Global studies can test whether the current effects are limited to the litigious US context or generalize to other geographics where AVs are being actively developed or deployed, like Europe and Asia, where we predict consumers will show the present response pattern so long as they view AVs as an abnormal presence that potentially interferes with human competence. Finally, even studies conducted within the United States can survey participants beyond the Prolific and Mturk samples studied here.

On interventions, future work can investigate how exposure to AVs, both by passengers of AVs and other drivers and pedestrians, affects the phenomena uncovered here. The question is whether exposure can serve to normalize AVs, thereby mitigating the bias found here, or whether negative public perceptions will be too persistent and a possible stopper for AV firms. Another approach may be to target consumer trust, by communicating that AVs do not only follow the literal rules of the road, but also take deliberate steps to evade accidents when they are not at fault. On this note, the current work studied the effect of trust in the AV's *competence*, given that competence is of primary concern for new technological products. But future work could also investigate whether other types of trust pertaining to the *AV firm* (rather than to the AV itself) influence liability, including trust based on the firm's benevolence (the extent to which the firm seems to want to do good to the trustor, regardless of profit incentive), and integrity (the extent to which the firm adheres to principles that the trustor thinks are reasonable) (Mayer et al., 1995; Sirdeshmukh et al., 2002; Xie & Peng, 2009).

### ACKNOWLEDGMENTS

For helpful comments, the authors thank John Deighton, Rajiv Lal, Elie Ofek, Rohit Desphande, Marc Freuler, Chris Moore, Rebecca Bredehoeft, and audience members of the 2022 New England Marketing Conference, and University of Chicago Booth's Marketing Unit and Roman Family Center for Decision Research.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available at: https://github.com/preacceptance/av_not_at_fault.

### DISCLOSURES

The results of this work were popularized in the Wall Street Journal: https://www.wsj.com/articles/will-we-blame-self-driving-cars-11674745636.

### ORCID

*Julian De Freitas* https://orcid.org/0000-0003-4912-1391

### REFERENCES

Arikan, E., Altinigne, N., Kuzgun, E., & Okan, M. (2023). May robots be held responsible for service failure and recovery? The role of robot service provider agents' human-likeness. *Journal of Retailing and Consumer Services*, *70*, 103175.

AUVSI. (2012). *Driverless car summit 2012: Conference report*. Association for Unmanned Vehicle Systems International.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*, 59–64.

Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2019). Drivers are blamed more than their automated cars when both make mistakes. *Nature Human Behaviour*, *4*, 1–10.

Bensinger, G. (2024). *Exclusive: GM's cruise valuation slashed by more than half, adding to woes*. Reuters.

Bloomberg. (2016). *Public perceptions of driverless cars*. Bloomberg Government.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*, 1573–1576.

Brennan, R., & Sachon, L. (2022). *Self-driving cars make 76% of Americans feel less safe on the road*. Policy Genius.

Cano, R. (2024). One crash set off a new era for self-driving cars in S.F. Here's a complete look at what happened.

Carlson, A. (2022). *You can now take a driverless Lyft in Austin. Here's what you need to know*. Austin American Statesman.

Castelo, N., & Ward, A. F. (2021). Conservatism predicts aversion to consequential Artificial Intelligence. *PLoS One*, *16*, e0261467.

Chafkin, M. (2022). Even after $100 billion, self-driving cars are going nowhere. *Businessweek*.

Darden, W. R., DeConinck, J. B., Babin, B. J., & Griffin, M. (1991). The role of consumer sympathy in product liability suits: An experimental investigation of loose coupling. *Journal of Business Research*, *22*, 65–89.

De Freitas, J., Anthony, S. E., Censi, A., & Alvarez, G. (2020). Doubting driverless dilemmas. *Perspectives on Psychological Science*, *15*, 1284–1288.

De Freitas, J., Censi, A., Smith, B. W., Di Lillo, L., Anthony, S. E., & Frazzoli, E. (2021). From driverless dilemmas to more practical commonsense tests for automated vehicles. *Proceedings of the National Academy of Sciences*, *118*, e2010202118.

De Freitas, J., & Cikara, M. (2021). Deliberately prejudiced self-driving cars elicit the most outrage. *Cognition*, *208*, 104555.

De Freitas, J., & Johnson, S. G. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology*, *79*, 149–163.

De Freitas, J., Ofek, E., Ingledew, S., & Labruyere, T. (2022). *Navya: Steering toward a driverless future*. Harvard Business Publishing.

Folkes, V., & Lassar, W. (2015). Counterfactuals and affective responses to product breakdown. In *Proceedings of the 1996 academy of marketing science (AMS) annual conference* (pp. 110–114). Springer.

Gill, T. (2020). Blame it on the self-driving car: How autonomous vehicles can alter consumer morality. *Journal of Consumer Research*, *47*, 272–291.

Gold, C., Körber, M., Hohenberger, C., Lechner, D., & Bengler, K. (2015). Trust in automation–before and after the experience of take-over scenarios in a highly automated vehicle. *Procedia Manufacturing*, *3*, 3025–3032.

Haboucha, C. J., Ishaq, R., & Shiftan, Y. (2017). User preferences regarding autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, *78*, 37–49.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*, 115–135.

HG.org. (2024). *My car is defective and caused an accident, what can I do?* HG.org.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*, 587–612.

Isidore, C., McFarland, M., & Valdes-Dapena, P. (2022). Ford takes $2.7 billion hit as it drops efforts to develop full self-driving cars.

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). Cambridge University Press.

Kalra, N., & Paddock, S. M. (2016). *Driving to safety: How many mile of driving would it take to demonstrate autonomous vehicle reliability*. RAND Corporation.

Kasimatis, M., & Wells, G. L. (2014). Individual differences in counterfactual thinking. In *What might have been* (pp. 93–114). Psychology Press.

Kolodny, L. (2022). *Cruise gets green light for commercial robotaxi service in San Francisco*. CNBC.

Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, *9*, 90–96.

Krishnamurthy, P., & Sivaraman, A. (2002). Counterfactual thinking and advertising responses. *Journal of Consumer Research*, *28*, 650–658.

Lavieri, P. S., Garikapati, V. M., Bhat, C. R., Pendyala, R. M., Astroza, S., & Dias, F. F. (2017). Modeling individual preferences for ownership and sharing of autonomous vehicle technologies. *Transportation Research Record*, *2665*, 1–10.

Lawyers, Q. E. T. (2024). *Report to the boards of directors of cruise LLC, GM cruise holdings LLC, and general motors holdings LLC regarding the October 2, 2023 accident in San Francisco (2024)*. Quinn, Emanuel, Urquhart, Sullivan, LLP.

Li, J., Zhao, X., Cho, M.-J., Ju, W., & Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. SAE Technical Paper.

Li, X., Hsee, C. K., & O'Brien, E. (2022). "It could Be better" can make it worse: When and why people mistakenly communicate upward counterfactual information. *Journal of Marketing Research*, *20*, 1312.

Longoni, C., Cian, L., & Kyung, E. J. (2022). Algorithmic transference: People overgeneralize failures of AI in the government. *Journal of Marketing Research*, *2*, 10139.

Loudenback, L. J., & Goebel, J. W. (1974). Marketing in the age of strict liability: The doctrine of strict liability has arrived. What is next for marketers? *Journal of Marketing*, *38*, 62–66.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*, 709–734.

Menon, N., Zhang, Y., Rawoof Pinjari, A., & Mannering, F. (2020). A statistical analysis of consumer perceptions towards automated vehicles and their intended adoption. *Transportation Planning and Technology*, *43*, 253–278.

Miller, D. T., & McFarland, C. (1986). Counterfactual thinking and victim compensation: A test of norm theory. *Personality and Social Psychology Bulletin*, *12*, 513–519.

Miller, D. T., Turnbull, W., & McFarland, C. (1990). Counterfactual thinking and social perception: Thinking about what might have been. In *Advances in experimental social psychology* (Vol. *23*, pp. 305–331). Elsevier.

Moorman, C., Zaltman, G., & Deshpande, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. *Journal of Marketing Research*, *29*, 314–328.

Morgan, F. W. (1982). Marketing and product liability: A review and update. *Journal of Marketing*, *46*, 69–78.

Morgan, F. W. (1987). Strict liability and the marketing of services vs. goods: A judicial review. *Journal of Public Policy & Marketing*, *6*, 43–57.

Moulton, S. H. (2019). *50-state analysis of liability damages caps*. US Law Network Inc.

Nader, R. (1965). *Unsafe at any speed: The designed-in dangers of the American automobile*. Grossman.

Nichols, C. (2013). *Liability could be roadblock for driverless cars*. The San Diego Union-Tribune.

Park, J., & Han, S. (2023). Investigating older consumers' acceptance factors of autonomous vehicles. *Journal of Retailing and Consumer Services*, *72*, 103241.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*, 4649–4654.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Rainie, L., Funk, C., Anderson, M., & Tyson, A. (2022). *Americans cautious about the deployment of driverless cars*. Pew Research Center.

Randazzo, R. (2020). *Are you ready to ride in a car without a driver/ Waymo vans going public in the Easy Valley*. Azcentral.

Rhee, E., Uleman, J. S., Lee, H. K., & Roman, R. J. (1995). Spontaneous self-descriptions and ethnic identities in individualistic and collectivistic cultures. *Journal of Personality and Social Psychology*, *69*, 142–152.

Roese, N. J., Summerville, A., & Fessel, F. (2007). Regret and behavior: Comment on Zeelenberg and Pieters. *Journal of Consumer Psychology*, *17*, 25–28.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*, 393–404.

Schoettle, B., & Sivak, M. (2014). *A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia*. University of Michigan, Ann Arbor, Transportation Research Institute.

Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2021). How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation Research Part C: Emerging Technologies*, *126*, 103069.

Singh, S. (2015). *Critical reasons for crashes investigated in the national motor vehicle crash causation survey*. National Highway Traffic Safety Administration.

Sirdeshmukh, D., Singh, J., & Sabol, B. (2002). Consumer trust, value, and loyalty in relational exchanges. *Journal of Marketing*, *66*, 15–37.

Smith, B. W. (2017). Automated driving and product liability. *Michigan State Law Review*, *2*, 1–74.

Smith, E. A. (2022). *Automated driving sensor and software recalls and lawsuits*. Autoaccident.com.

Srinivasan, R., & Sarial-Abi, G. (2021). When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing*, *85*, 74–91.

Strahilevitz, M. A., Odean, T., & Barber, B. M. (2011). Once burned, twice shy: How naive learning, counterfactuals, and regret affect the repurchase of stocks previously sold. *Journal of Marketing Research*, *48*, S102–S120.

SwissRe. (2022). *Driver today, passenger next*. Swiss Re.

Tsiros, M., & Mittal, V. (2000). Regret: A model of its antecedents and consequences in consumer decision making. *Journal of Consumer Research*, *26*, 401–417.

Villasenor, J. (2014). *Products liability and driverless cars: Issues and guiding principles for legislation*. Brookings.

Vynck, G. D. (2022). *The tech CEO spending millions to stop Elon musk*. The Washington Post.

Welle, B., Sharpin, A. B., Adriazola-Steil, C., Job, S., Shotten, M., Bose, D., Bhatt, A., Alveano, S., Obelheiro, M., & Imamoglu, T. (2018). *Sustainable & safe: A vision and guidance for zero road deaths*. World Resources Institute.

Wells, K. (2022). *Swiss Re collaborates with Waymo to study autonomous vehicles*. Reinsurance News.

Wessling, B. (2022). *Motional, Lyft begin autonomous rides in Las Vegas*. The Robot Report.

Wiggin, K. L., & Yalch, R. F. (2015). Whose fault is it? Effects of relational self-views and outcome counterfactuals on self-serving attribution biases following brand policy changes. *Journal of Consumer Psychology*, *25*, 459–472.

Wright, R. W. (1992). The logic and fairness of joint and several liability. *Mem. St. UL Rev.*, 45.

Xie, Y., & Peng, S. (2009). How to repair customer trust after negative publicity: The roles of competence, integrity, benevolence, and forgiveness. *Psychology & Marketing*, *26*, 572–589.

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. *Transportation Research Part C: Emerging Technologies*, *95*, 320–334.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** De Freitas, J., Zhou, X., Atzei, M., Boardman, S., & Lillo, L. D. (2025). Public perception and autonomous vehicle liability. *Journal of Consumer Psychology*, *00*, 1–16. https://doi.org/10.1002/jcpy.1448