Consumers Hold Autonomous Vehicles Liable Even When Not at Fault

Julian De Freitas
Xilin Zhou
Margherita Atzei
Shoshana Boardman
Luigi Di Lillo
Consumers Hold Autonomous Vehicles Liable Even When Not at Fault

Julian De Freitas
Harvard Business School

Xilin Zhou
Swiss Reinsurance Company, Ltd.

Margherita Atzei
Swiss Reinsurance Company, Ltd.

Shoshana Boardman
Harvard University

Luigi Di Lillo
Swiss Reinsurance Company, Ltd.

Working Paper 23-036

Copyright © 2023, 2024 by Julian De Freitas, Xilin Zhou, Margherita Atzei, Shoshana Boardman, and Luigi Di Lillo.
Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only. It may not be reproduced without permission of the copyright holder. Copies of working papers are available from the author.
Funding for this research was provided in part by Harvard Business School.
Consumers Hold Autonomous Vehicles Liable Even When Not at Fault

Abstract (171 Words)

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 four experiments (N=4,061), 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 entertain counterfactuals in which the not-at-fault vehicle might have behaved differently to avoid, or minimize damage from, the accident. Given this reasoning, consumers conclude that in such a case, the vehicle could have acted more optimally to prevent or avoid the accident, even if it did not cause it, leading 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; moral judgment; insurance; liability; harm
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 and Wagner 2017). They could reduce congestion by driving optimally; free human attention to converse, conduct meetings, drive tired, or just sleep; 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 drivers¹.

Yet, 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 USA, 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; Smith 2022; Villasenor 2014). Most notably, Cruise, a subsidiary of General Motors, recently lost its license to operate in San Francisco after one of its autonomous vehicles was involved in an accident where it was not at fault (Bensinger 2024; Cano 2024).

¹ 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).
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. While not-at-fault accidents are not typically considered liability risks for human drivers or human-driven vehicle (HDV) manufacturers, we ask whether the public thinks the firm which 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 conceptual background and our theoretical framework, followed by four studies (which are buttressed by four further supplemental studies) that test the proposed response pattern and underlying counterfactual thought process. 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 and 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, e.g., driving a tire at exceedingly high speeds.

Thus far, the automative 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. 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’, and the manufacturer 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 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, which are pertinent for several reasons. First, consumers will be the plaintiffs in lawsuits against AV firms, which are most likely to go to trial when victims are seriously injured. 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. 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).

---

3 https://www.law.cornell.edu/wex/class_action
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, because AVs are not yet widely deployed, there is a dearth of claims data available (Wells 2022), making it difficult to apply traditional actuarial approaches for risk-assessment (SwissRe 2022). Additionally, since AVs have not yet driven sufficient miles to afford a statistically meaningful safety comparison with HDVs (Kalra and Paddock 2016), manufacturers and insurers must turn to alternative approaches to estimate and explain liability risk.

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 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 and 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 and Sachon 2022). Many hesitations stem from safety concerns over the performance and failures of automated vehicles, as well as fear of ceding control to a machine (Schoettle and Sivak 2014; Shariff, Bonnefon, and Rahwan 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; De Freitas et al. 2021; De Freitas and Cikara 2021).

Much of this work finds that AV adoption boils down to whether consumers trust in 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 this poses 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, Ishaq, and Shiftan 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 role of counterfactual simulation. With that said, we do operationalize trust as an individual difference variable that may impact this counterfactual reasoning process. 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 and 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).

**Counterfactual Simulation**

Next, our proposed mechanism draws on work on counterfactual simulation. Consumer judgments are sometimes affected by counterfactuals (Folkes and Lassar 2015; Tsiros and Mittal 2000; Wiggin and Yalch 2015)—psychological simulations of how events could have turned out differently, had an alternative course of action been taken (Kahneman and 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 one 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 and Tversky 1982).

More broadly, consumers tend to think of counterfactuals when an event is ‘abnormal’, deviating from the statistical or social norm (Hitchcock and Knobe 2009; Miller and McFarland
1986; Phillips, Luguri, and Knobe 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, Turnbull, and McFarland 1990; Wiggin and Yalch 2015).

Within consumer psychology, counterfactual simulation has been implicated in a few notable domains, including: post-purchase regret and consumption choices (Roese, Summerville, and Fessl 2007; Strahilevitz, Odean, and Barber 2011; Tsiros and Mittal 2000); responses to product breakdown and brand transgressions (Folkes and Lassar 2015; Wiggin and Yalch 2015); and promotion tactics and consumer responses to them (Krishnamurthy & Sivaraman, 2002; Li, Hsee, & O’Brien, 2022. In the current work, we explore whether and how 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 and 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 and Johnson 2018; Phillips and Cushman 2017). Notably, they expect agents to behave optimally even when it is unfair to do so, 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 and Johnson 2018). The current work similarly tests whether, when participants imagine counterfactuals in which the driver of the not-at-fault vehicle was a human rather than an AV, they imagine 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 ensured by either the design of the scenario (all Studies except Study 2)
and/or affirmed by experts (Study 2).

**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.

We expect that when the accident occurs, consumers begin to search for a causal
explanation (Kahneman and 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, thus minimizing the 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
entertain a counterfactual scenario where the human occupant was in control and could have
potentially avoided the accident. Because participants are inclined to imagine counterfactuals in
which an agent acts optimally, they conclude that a human in the same position could have
avoided the accident, leading them to assign liability to the AV manufacturer for not designing
the vehicle to behave the same. Given that this proposed mechanism separates whether the
counterfactual seems relevant (aka counterfactual relevance) from the evaluation stemming from
this counterfactual (aka avoidance conclusion), we measure each of these as separate
psychological constructs.

We also expect that liability ascriptions are affected by how constrained the accident
scenario itself was. The more constrained the scenario is, the fewer counterfactuals there are for
how the vehicle could have behaved differently, and the less liable the AV can be viewed. We expect that the scenario constraints do not necessarily impact whether a counterfactual seems relevant in the first place, which should be immediately influenced by the presence of an ‘abnormal’ AV. However, we do expect that the constraints impact the specific counterfactuals people can generate when they go on to do so, as when they must make a judgment about whether the AV could have avoided the accident—Figure 1.

Finally, 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, reducing consumers’ tendency to engage in counterfactual thinking around what the AV could have done differently.

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 toward AVs moderate the effects of counterfactual relevance on liability judgments. This model provides a new theoretical lens for understanding the interplay between technology perception and legally relevant judgments.

<<<INSERT FIGURE 1 HERE>>> 

**Overview of Studies**

We ask whether AV firms are viewed as more liable than HDV firms for accidents not at fault, whether this response pattern is driven by the proposed counterfactual mechanism (Studies 1-4), and whether the effect of vehicle type on counterfactual simulation is moderated by trust in AVs (Studies 1-2). In addition to replicating our effects around various realistic hypothetical scenarios, we replicate it in a scenario modeled directly off of a real autonomous vehicle accident
(Study 2). Finally, we show that the effect is moderated by whether the scenario is constrained (Study 3), or whether the at-fault vehicle’s traffic violating behavior is highlighted (Study 4). 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: [anonymized repo link].

Altogether, we find robust evidence for all our conclusions from a total of 4,061 participants. We also include a pre-study and three conceptual replications of Study 1, in the MDA. Including these studies, we present evidence from 6,314 participants. For generalization purposes, we sample participants from both the Mturk and Prolific platforms, and exclusion percentages never exceeded 12%, apart from Study 2 which additionally excluded 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 non-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 MDA). 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 and participants are more likely to consider the counterfactual in which the not-at-fault vehicle might have done something differently to avoid, or minimize damage from, 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. We predict that this difference will also be mediated by the proposed counterfactual reasoning process.

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

**Method**

This study was pre-registered ([https://aspredicted.org/DHL_WV9](https://aspredicted.org/DHL_WV9)), and we did not deviate from the pre-registered plan. We recruited 904 participants from Amazon’s Mechanical
Turk, who passed attention checks and completed the survey, in exchange for $0.50. We excluded 104 for failing the same comprehension questions as Study 1, leaving 800 participants ($M_{age} = 41.6, 55.9\%$ females). Participants were only allowed to participate if they correctly answered two attention checks at the beginning of the survey. Participants were evenly assigned to one of two vehicle type conditions (HDV or AV) and given the following instruction (information in squared parentheses was only included in the AV condition), accompanied by Figure 2:

*You will watch an animated video of a traffic scenario, depicted below. The video shows a four-way 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.]*

<<<INSERT FIGURE 2 HERE>>> Readers can view the videos here: [https://youtu.be/3UQ1GkjTZk0](https://youtu.be/3UQ1GkjTZk0) (HDV); [https://youtu.be/J9zzSe-VHiM](https://youtu.be/J9zzSe-VHiM) (AV). Participants were required to watch the video twice, then indicate 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 relevant AV or HDV still image from Figure 2. The dependent variables pertained to the liability of the at-fault and not-at-fault vehicles, although we were chiefly interested in the latter (Table 1).

<<<INSERT TABLE 1 HERE>>>
First, the DV measures were presented in randomized order. Next, the counterfactual mediator measures were presented also in randomized order. Finally, the measure of Vehicle B’s ability to do more was presented. 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 ran the stop sign (vehicle A or B).

Next, participants 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 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. Finally, participants completed two comprehension items and standard demographics questions. Participants who failed either of the comprehension checks were excluded from 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
Unsurprisingly, the at-fault vehicle (vehicle A in the scenario) was viewed as highly liable whether it struck an AV or HDV.

Crucially, however, the manufacturer of the not-at-fault vehicle (vehicle B in the scenario) was viewed as more liable when its vehicle was an AV versus HDV ($M_{AV} = 20.95$ $M_{HDV} = 11.60$, $t(771) = 5.49$, $p < .001$, $d = 0.38$). In addition, participants thought the AV manufacturer was more liable than a human driver of an HDV when both were not-at-fault ($M_{AV} = 20.95$, $M_{HDV} = 10.94$, $t(797) = 5.51$, $p < .001$, $d = 0.39$). 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.

These results may also explain why participants thought it was less relevant to consider the counterfactual for the at-fault driver, and ultimately viewed it as less liable, when it struck an AV as opposed to an HDV. The AV may have distracted attention away from the at-fault vehicle, such that participants who considered counterfactuals for the not-at-fault AV were less likely to also do so for the at-fault HDV. Thus, due to the same underlying counterfactual mechanism implicated in our main effect, moral judgments tend to be zero sum rather than cumulative—even though in court multiple parties can technically be held liable (De Freitas & Hafri, 2024; Gray & Wegner, 2009).

In line with our hypothesized thought process, we also found that participants were more likely to consider the counterfactual (in which Vehicle B had behaved differently) when the not-at-fault vehicle was an AV than an HDV, and they were more likely to conclude that the not-at-fault vehicle could have avoided the accident (Table 2).
As for trust in AVs, we used a single measure of AV trust averaged across all five measures ($\alpha = 0.93$). We found that participants were more likely to trust AVs when the not-at-fault vehicle was an AV (Table 2). The difference indicates that participants’ trust in AVs was likely driven by the manipulation itself.

Mediation analysis. We conducted a serial mediation analysis (PROCESS Model 6; Hayes 2012) to determine whether the effect of vehicle type on manufacturer liability was serially mediated in the following hypothesized causal order: condition $\rightarrow$ counterfactual $\rightarrow$ could have done more $\rightarrow$ DV. The serial mediation was statistically significant for the manufacturer comparison ($b = -2.11$, $SE = 0.58$, 95% CI [-3.35, -1.09]).

Our mediation output also included the results for two simpler parallel mediation models. The indirect effect, condition $\rightarrow$ could have done more $\rightarrow$ DV, was not statistically significant ($b = 0.01$, $SE = 0.34$, 95% CI [-0.71, 0.67]). But the model, condition $\rightarrow$ counterfactual $\rightarrow$ DV, was significant ($b = -1.22$, $SE = 0.50$, 95% CI [-2.32, -0.37]). However, we continued to favor the serial model because (i) the simpler parallel model did not differ from our proposed serial model ($b_{\text{difference}} = 0.89$, $SE = 0.77$, 95% CI [-0.55, 2.47]), (ii) the serial model showed a numerically stronger effect ($b_{\text{parallel}} = -1.22$ vs. $b_{\text{serial}} = -2.11$), and (iii) given our theoretical reasons for testing the serial model.

Next, we conducted a moderated serial mediation analysis (PROCESS Model 83; Hayes 2012), testing for the same serial model but with the path between vehicle condition and counterfactual moderated by trust in AVs. We found that the index of moderated mediation was significant for the manufacturer comparison ($b = 0.08$, $SE = 0.02$, 95% CI [0.04, 0.12]; Figure 4).
In line with our predictions, the less that participants trusted AVs, the more likely they were to entertain the counterfactual (Figure 5).

We additionally tested an alternative hypothesis where the effect of vehicle type on manufacturer liability is instead mediated through trust in AVs. We conducted a parallel mediation analysis (PROCESS Model 4; Hayes 2012) testing for the following path: condition \( \rightarrow \) trust in AVs \( \rightarrow \) DV. We found this alternative mechanism was not statistically significant \( (b = 0.28, SE = 0.27, 95\% CI [-0.21, 0.85]) \), indicating that the manipulation’s effect on AV trust does not itself explain the main findings.

For all four 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 and Han 2023). Using simple linear regression, we do not find statistically significant moderation effects for age across the four studies.

Across all four studies, we additionally ensure discriminant validity between measures by testing and confirming that correlations among all measures were less than 0.80 (see MDA). We note that some methods for assessing discriminant validity, such as Fornell-Larcker criterion, require multi-item constructs where each construct is represented by multiple observed variables. Given the single-item nature of most of our constructs, these methods are not directly applicable to our dataset.

\<<<INSERT FIGURE 4 HERE>>>
\<<<INSERT FIGURE 5 HERE>>>

**Discussion**

Consumers viewed AV manufacturers as more liable than HDV manufacturers or human drivers for accidents not at fault. This discrepancy was rooted in a greater tendency to entertain
counterfactual scenarios in which the not-at-fault vehicle had behaved differently when the vehicle was an AV. Given these mental simulations, participants concluded that the AV could have done more, hence their assessments that AV manufacturers were more liable. Additionally, we found that the degree to which consumers generated counterfactual simulations more so 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 MDA we report three conceptual replications of these results. The first replication study shows similar results for a downstream measure of willingness to sue the company and shows that mere driving expectations for AVs does not explain the effects. It also 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 second replicates the results in a different driving scenario. The third replicates the results in a third driving scenario, and tests 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 and Ward 2021), would be more inclined to consider counterfactuals 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

As a stronger test of the ecological validity of our findings, Study 2 tests 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 that did not brake, then thrown into the path of a Cruise vehicle. Despite the Cruise vehicle’s attempt to brake (emanuel 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 Study 1 is whether Cruise was penalized more heavily merely because its vehicle was autonomous.

Study 2 also makes several other additional contributions. First, it includes a different, more downstream measure, of whether consumers would be willing to sue the firm. Second, it removes the wifi symbol that was used to depict the AV in the videos of Study 1, which might have influenced responses in unanticipated ways. Third, it includes the trust measure at the beginning of the experiment, rather than the end, so that it cannot be influenced by the condition manipulation.

Method

This study was pre-registered (https://aspredicted.org/3Q2_GYD), and we did not deviate from the pre-registered plan. We recruited 895 participants 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 assigned to one of two conditions (agent: AV or HDV) in a between-subjects design. Participants first answered the same trust questions from Study 1. 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 6.

Participants then answered the same questions from Study 1, accompanied by the schematics in Figure 6, except that the prior questions about liability were substituted with a more downstream rating of willingness to sue: “How much do you agree: It would be reasonable to sue the [driver of vehicle A / manufacturer of Vehicle B] to cover the costs of the serious injuries sustained by the pedestrian”. Then, participants completed two comprehension checks, 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 analysis.

Results

We successfully replicated all significant differences found in Study 1 (Table 3; Figure 7). Participants once again viewed the manufacturer of a not-at-fault vehicle as more liable when the not-at-fault vehicle was an AV. Specifically, participants were more inclined to sue the manufacturer of a not-at-fault AV than either the manufacturer ($M_{AV} = 38.33$ $M_{HDV} = 30.63$, $t(716) = 3.24,$ $p = .001, d = 0.24$) or driver ($M_{AV} = 38.33$ $M_{HDV} = 14.08$, $t(642) = 11.63,$ $p < .001, d = 0.87$) of the same vehicle had it instead been a conventional HDV. In addition, participants were again highly willing 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.
In line with the hypothesized thought process established in Study 1, we again found that participants were more likely to consider the counterfactual (in which Vehicle B had behaved differently) when Vehicle B was an AV than an HDV, and they were more likely to conclude that Vehicle B could have avoided the accident (Table 3).

The AV trust measure was once again averaged across all five trust measures ($\alpha = 0.92$). In contrast to Study 1, and as intended given that this time trust was presented at the beginning of the study, there was no difference in trust between conditions (Table 3).

**Mediation analysis.** We successfully replicated the serial and moderated serial mediation findings established in Study 1. We first conducted a serial mediation analysis (PROCESS Model 6; Hayes 2012) to determine whether the hypothesized causal order: condition $\rightarrow$ counterfactual $\rightarrow$ could have done more $\rightarrow$ DV, extended to the more ecologically valid collision scenario. The serial mediation was statistically significant for the manufacturer comparison ($b = -1.63, SE = 0.84, 95\% CI [-3.33, -0.04]$).

Again, our mediation output also included two simpler parallel mediation models. As in Study 1, the parallel model with “could have done more” was not significant ($b = -1.00, SE = 0.78, 95\% CI [-2.57, 0.51]$), whereas parallel model with “counterfactual” as the mediator was significant ($b = -1.03, SE = 0.56, 95\% CI [-2.21, -0.03]$). Again, this simpler model did not differ significantly from our proposed serial model ($b_{\text{difference}} = 0.60, SE = 0.49, 95\% CI [-0.11, 1.77]$) and exhibited a numerically weaker effect than the serial model ($b_{\text{parallel}} = -1.03, b_{\text{serial}} = -1.63$).

Next, we tested the same moderated serial mediation analysis as study 1 (PROCESS Model 83; Hayes 2012), in which the path between vehicle condition and counterfactual is moderated by baseline trust in AVs. The index of moderated mediation was significant for the
manufacturer comparison ($b = 0.15, \ SE = 0.04, \ 95\%\ CI [0.08, 0.23]; \ Figure\ 8$). The less participants trusted AVs at baseline, the more they considered the counterfactual relevant.

<<<INSERT FIGURE 8 HERE>>>  

**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 counterfactual 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 the 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 avoided the accident.

**Study 3**

In order to causally test the validity of the counterfactual mechanism, Study 3 investigates whether we can turn off or significantly weaken the main effect by manipulating the availability of alternate driving paths that the not-at-fault vehicle could take. We anticipate that if there are fewer counterfactual scenarios available in which a not-at-fault vehicle could have avoided the accident, then the tendency to think that manufacturers of AVs are more liable is diminished or eliminated altogether. Specifically, we predict that whether the scenario is constrained or not moderates the path between counterfactual relevance and judgments that the not-at-fault vehicle could have avoided the accident, because more constrained scenarios inhibit participants from thinking of specific counterfactuals that could implicate the vehicle.
Method

This study was pre-registered (https://aspredicted.org/CTL_8RL), and we did not deviate from the pre-registered plan. We recruited 902 participants from Amazon’s Mechanical Turk, who passed attention checks and completed the survey, in exchange for $0.50. We excluded 102 for failing similar comprehension checks as Studies 1-3 (described below), yielding 800 participants (M_age = 40.1, 60.1% females).

Participants were assigned to one of four conditions in a 2 (agent: AV or HDV) x 2 (scenario: constrained or unconstrained) between-subjects design. In the unconstrained condition, the not-at-fault Vehicle B was shown driving down a one-way street. In the constrained condition, the scenario was identical except the degrees of freedom of Vehicle B were severely constrained by parked cars on either side of the path and a tailing vehicle. Participants in all conditions were given the following instructions (information in squared parentheses was only included in the AV condition) accompanied by Figure 9:

*You will watch an animated video of a traffic scenario, depicted below. The video depicts a one-way street in which the driver of Vehicle A is driving down the wrong way 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.]*

<<<INSERT FIGURE 9 HERE>>>

Readers can view the videos here: [http://y2u.be/mDd_m0hSq24](http://y2u.be/mDd_m0hSq24) (HDV, Unconstrained); [http://y2u.be/PHk0-RczBWC](http://y2u.be/PHk0-RczBWC) (AV, Unconstrained); [http://y2u.be/pM0093feCoY](http://y2u.be/pM0093feCoY) (HDV, Constrained); [http://y2u.be/GtIrXW9pe0](http://y2u.be/GtIrXW9pe0) (AV, Constrained). Participants watched the assigned video twice. After completing the same main measures and checks as in Studies 1-3, participants
were asked to answer two comprehension checks about what type of vehicle they saw in the scenario (AV or HDV) and which vehicle was driving the wrong way (vehicle A or B).

**Results**

We ran 2 (agent: AV or HDV) x 2 (scenario: constrained or unconstrained) 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 ($M_{AV} = 20.36, M_{HDV} = 8.10, F(1, 796) = 53.81, p < .001, \eta^2 = 0.063$), scenario type ($M_{unconstrained} = 18.17, M_{constrained} = 10.82, F(1, 796) = 17.71, p < .001, \eta^2 = 0.022$), and an interaction effect ($F(1, 796) = 8.79, p = .003, \eta^2 = 0.011$). Participants were more likely to judge the AV manufacturer as liable than the HDV manufacturer, both in the unconstrained condition ($M_{AV} = 26.29, M_{HDV} = 9.12, t(346) = 6.61, p < .001, d = 0.66$) and constrained condition ($M_{AV} = 14.45, M_{HDV} = 7.19, t(348) = 3.53, p < .001, d = 0.35$), although the effect was much larger in the unconstrained condition.

For AV manufacturer vs human driver liability, we found a main effect of agent type ($M_{AV} = 20.36, M_{HDV} = 5.96, F(1, 796) = 73.33, p < .001, \eta^2 = 0.084$), scenario type ($M_{unconstrained} = 17.63, M_{constrained} = 9.30, F(1, 796) = 22.44, p < .001, \eta^2 = 0.027$), and an interaction effect ($F(1, 796) = 5.63, p = .018, \eta^2 = 0.007$). Participants were more likely to judge the AV manufacturer as liable than the human driver, both in the unconstrained condition ($M_{AV} = 26.29, M_{HDV} = 7.99, t(358) = 6.89, p < .001, d = 0.69$) and constrained condition ($M_{AV} = 14.45, M_{HDV} = 4.13, t(339) = 5.08, p < .001, d = 0.50$), although the effect was larger in the unconstrained condition.

<<<INSERT TABLE 4 HERE>>>
relevance and judgments that the not-at-fault vehicle could have done more to avoid the accident. This test deviates from our pre-registered plan to test moderation of the path from agent condition to counterfactual relevance, since we realized that actual counterfactual simulations are most likely to occur when judging whether the vehicle could have avoided the accident.

We conducted a moderated serial mediation analysis (PROCESS Model 4; Hayes 2012) in which the middle path of the serial model is moderated by constrained condition, thereby affecting judgments of liability for the vehicle manufacturer. The index of moderated mediation was significant ($b = -0.54, SE = 0.24, 95\% \text{ CI} [-1.10, -0.16])$.

**Discussion**

We found that manipulating the availability of counterfactual driving paths for a not-at-fault vehicle significantly attenuates the liability risk for AV firms (vs. HDV firms and human drivers), by constraining what counterfactuals inform a judgment of whether the not-at-fault vehicle could have done more to avoid the accident. This result both strengthens evidence for the proposed counterfactual mechanism and suggests that interventions for decreasing liability risk should focus on constraining the scenario. For instance, defendants of firms might make arguments for why certain counterfactuals are not feasible.

At the same time, we continued to find a significantly higher liability risk for AV firms in the constrained scenarios, which is consistent with two interpretations: First, consumers may still generate counterfactuals, but at a broader scope that extends beyond the specifics of the scenario itself—for example, they may imagine the AV not taking the path in the first place. Second, they may have negative baseline views toward AV firms that induce them to ascribe higher liability even when they cannot imagine how the vehicle could have behaved otherwise. In this way, lower average trust of these vehicles might separately contribute to liability risk, in addition to
the counterfactual thought process; although we found no evidence for this in the prior studies.

Study 4

Study 4 investigates whether we can intervene on the main effect by deflecting the participant’s attention away from the not-at-fault vehicle and focusing 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 towards the abnormality of the AV’s presence. Therefore, we anticipate that if participants are provided additional details on the at-fault vehicle’s violations, then they will be less likely to consider counterfactuals involving the not-at-fault AV, thereby diminishing or eliminating their tendency to view not-at-fault AV manufacturers as more liable. Specifically, we predict that highlighting the faults of the at-fault vehicle moderates the path between vehicle type and counterfactual relevance. To ensure further robustness, we also use another driving scenario.

Methods

This study was pre-registered (https://aspredicted.org/NJ1_X1D), and we did not deviate from the pre-registered plan. We recruited 1360 participants from Amazon’s Mechanical Turk, who passed attention checks and completed the survey, in exchange for $0.70. We excluded 121 for failing similar comprehension checks as Studies 1-3 (described below), yielding 1239 participants ($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 10:
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 10.

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 Studies 1-3, 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) x 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 ($M_{AV} = 14.59$, $M_{HDV} = 10.40$, $F(1, 1235) = 11.89$, $p < .001$, $\eta^2 = 0.010$), no main effect of intervention ($M_{intervention} =
11.82, $M_{\text{no intervention}} = 13.26, \ F(1, 1235) = 1.42, \ p = .234, \ \eta^2 = 0.001$), and an interaction effect ($F(1, 1235) = 7.94, \ p = .005, \ \eta^2 = 0.006$). As in prior studies, participants were more likely to judge the AV manufacturer as liable than the HDV manufacturer without the intervention ($M_{AV} = 17.00, \ M_{HDV} = 9.38, \ t(554) = 4.43, \ p < .001, \ d = 0.35$). However, when participants received additional details on the at-fault vehicle’s traffic violations, there was no longer a significant difference between AV and HDV manufacturer liability for the not-at-fault vehicle ($M_{AV} = 12.20, \ M_{HDV} = 11.43, \ t(617) = 0.45, \ p = .651, \ d = 0.04$). The intervention successfully eliminates the main effect found in Studies 1-3.

For AV manufacturer vs human driver liability, we found a main effect of agent type ($M_{AV} = 14.59, \ M_{HDV} = 6.36, \ F(1, 1235) = 43.16, \ p < .001, \ \eta^2 = 0.034$), no main effect of intervention ($M_{\text{intervention}} = 9.60, \ M_{\text{no intervention}} = 11.52, \ F(1, 1235) = 2.41, \ p = .121, \ \eta^2 = 0.002$), and an interaction effect ($F(1, 1235) = 5.42, \ p = .020, \ \eta^2 = 0.004$). Participants were more likely to judge the AV manufacturer as liable than the human driver, both without the intervention ($M_{AV} = 17.00, \ M_{HDV} = 5.85, \ t(591) = 6.18, \ p < .001, \ d = 0.49$) and with the intervention ($M_{AV} = 12.20, \ M_{HDV} = 6.88, \ t(615) = 3.08, \ p = .002, \ d = 0.25$), although the effect was smaller with the intervention.

We additionally ran an ANOVA for our measure of counterfactual relevance, finding a main effect of agent type ($M_{AV} = 29.23, \ M_{HDV} = 19.29, \ F(1, 1235) = 36.11, \ p < .001, \ \eta^2 = 0.028$), no main effect of intervention ($M_{\text{intervention}} = 24.10, \ M_{\text{no intervention}} = 24.63, \ F(1, 1235) = 0.11, \ p = 0.737, \ \eta^2 = 0.000$), and an interaction effect ($F(1, 1235) = 5.52, \ p = .019, \ \eta^2 = 0.004$). Participants were more likely to rate the counterfactual as relevant when the not-at-fault vehicle was an AV, both without the intervention ($M_{AV} = 31.42, \ M_{HDV} = 17.59, \ t(578) = 6.16, \ p < .001, \ d = 0.49$) 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.

<<<INSERT TABLE 5 HERE>>>

**Mediation Results.** We tested whether the intervention condition moderates the A path of
the serial model between vehicle type condition and counterfactual relevance (PROCESS Model
83; Hayes 2012). The index of moderated mediation was found to be significant for the
manufacturer vs manufacturer comparison, ($b = -0.75, SE = 0.36, 95\% \text{ CI } [-1.51, -0.12]$). 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). Specifically, the intervention reduced the relevance of entertaining
counterfactuals where the AV or a human driver in the AV’s position could have behaved
differently. This result both strengthens 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 four studies and four supplemental studies, we found that vehicle manufacturers
are more likely to be viewed as liable when their vehicles are autonomous than 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 a human had more control over the AV, and conclude that a human would have done more to avoid the accident; hence, an AV firm was more liable than an HDV firm in the same scenario. Similarly, the AV firm was seen as more liable than the human driver of a not-at-fault HDV, showing that in practice it was held to a higher standard than all comparable parties. Liability 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 counterfactuals in which the AV acted differently. Likewise, liability 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 finds lower blame for AI than humans, driven by inferences of lower agency for AI (Arikan et al. 2023; Li et al. 2016; Srinivasan and Sarial-Abi 2021). We find the reverse effect in scenarios where AI is not at fault, due to a distinct counterfactual mechanism in which ‘abnormal’ AVs trigger the 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, Shariff, and Rahwan 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 think of 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 come to mind in event-based scenarios, and that this influences inferences about product defects and firm liability.

**Practical Implications**

If consumers are inclined to hold AV firms more liable 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 ridesharing 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 estimated cost of damages as a form of punishment⁴. 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 high frequency of such accidents, 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, McFarland, and Valdes-Dapena 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. Another silver lining is that autonomous vehicles that operate within a more constrained circuit, as in some automated buses (De Freitas et al. 2022), may be less susceptible to litigation because consumers may be less likely to imagine how they could have counterfactually avoided accidents.

Finally, AV manufacturers can also reduce 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 AV.

**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, Cian, and Kyung 2022). If such an inference is at play, it would be in addition to the counterfactual mechanism uncovered here, which was causally implicated in the response pattern. Future work could also further probe
whether negative attitudes toward AVs, such as believing that they will eliminate human jobs, affect liability ascriptions in addition to the counterfactual 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 U.S. 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 U.S. 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 but also by 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, Davis, and Schoorman 1995; Sirdeshmukh, Singh, and Sabol 2002; Xie and Peng 2009).

References


Bensinger, G (2024), "Exclusive: Gm's Cruise Valuation Slashed by More Than Half, Adding to Woes."


Castelo, Noah and Adrian F Ward (2021), "Conservatism Predicts Aversion to Consequential Artificial Intelligence," *PloS one*, 16 (12), e0261467.


35

De Freitas, Julian, Andrea Censi, Bryant Walker Smith, Luigi Di Lillo, Sam E Anthony, and Emilio Frazzoli (2021), "From Driverless Dilemmas to More Practical Commonsense Tests for Automated Vehicles," Proceedings of the National Academy of Sciences, 118 (11), e2010202118.

De Freitas, Julian and Mina Cikara (2021), "Deliberately Prejudiced Self-Driving Cars Elicit the Most Outrage," Cognition, 208, 104555.


emanuel, quinn (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," https://assets.ctfassets.net/95kuvdy8zn1v/1mb55pLYkkXVn0nXxEXz7w/9fb0e4938a89dc5ce09bf39e86ce5b9c/2024.01.24_Quinn_Emanuel_Report_re_Cruise.pdf.


Table 1.
Measures in Study 1.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liability, At Fault (DV)</td>
<td>The driver of Vehicle A is liable for the serious injuries sustained by the occupant of Vehicle B.</td>
</tr>
<tr>
<td>Liability of manufacturer, Not at Fault (DV)</td>
<td>The manufacturer of Vehicle B is liable for the serious injuries sustained by the occupant of Vehicle B.</td>
</tr>
<tr>
<td>Liability of driver, Not at Fault* (DV)</td>
<td>The driver of Vehicle B is liable for the serious injuries sustained by the occupant of Vehicle B.</td>
</tr>
<tr>
<td>Counterfactual, At Fault (M)</td>
<td>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.</td>
</tr>
</tbody>
</table>
Counterfactual, 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.

Notes: (1) * only measured in HDV condition; (2) DV=dependent variable, M=mediator, MOD=moderator.

Table 2.  
Statistics for Study 1.

<table>
<thead>
<tr>
<th>Measure</th>
<th>AV Mean</th>
<th>HDV Mean</th>
<th>T-Value</th>
<th>Cohen’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liability, At Fault (DV)</td>
<td>93.72 (14.07)</td>
<td>96.30 (9.23)</td>
<td>t(734)=3.09**</td>
<td>-0.21</td>
</tr>
<tr>
<td>Firm Liability, Not at Fault (DV)</td>
<td>20.95 (27.52)</td>
<td>11.60 (20.39)</td>
<td>t(771)=5.49***</td>
<td>0.38</td>
</tr>
<tr>
<td>Counterfactual, At Fault (M)</td>
<td>93.61 (15.43)</td>
<td>96.32 (12.11)</td>
<td>t(785)=2.78**</td>
<td>-0.19</td>
</tr>
<tr>
<td>Counterfactual, Not at Fault (M)</td>
<td>44.30 (34.26)</td>
<td>32.65 (31.88)</td>
<td>t(797)=4.98***</td>
<td>0.35</td>
</tr>
<tr>
<td>Done more, Not at Fault (M)</td>
<td>38.68 (32.80)</td>
<td>29.85 (30.97)</td>
<td>t(796)=3.91***</td>
<td>0.28</td>
</tr>
<tr>
<td>Trust in Autonomous Vehicles</td>
<td>38.39 (26.60)</td>
<td>30.94 (25.65)</td>
<td>t(794)=4.03***</td>
<td>0.29</td>
</tr>
</tbody>
</table>

Note: T-statistics reflect results of independent-samples t-tests. ** p < .01, *** p < .001.

Table 3.  
Statistics for Study 2.

<table>
<thead>
<tr>
<th>Measure</th>
<th>AV Mean</th>
<th>HDV Mean</th>
<th>T-Value</th>
<th>Cohen’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sue, Vehicle A (DV)</td>
<td>87.60 (19.93)</td>
<td>81.14 (26.48)</td>
<td>t(661)=3.69***</td>
<td>0.28</td>
</tr>
<tr>
<td>Firm Sue, Vehicle B (DV)</td>
<td>38.33 (32.57)</td>
<td>30.63 (31.09)</td>
<td>t(716)=3.24**</td>
<td>0.24</td>
</tr>
<tr>
<td>Counterfactual, Vehicle A (M)</td>
<td>91.73 (14.81)</td>
<td>87.29 (19.49)</td>
<td>t(664)=3.43***</td>
<td>0.26</td>
</tr>
<tr>
<td>Counterfactual, Vehicle B (M)</td>
<td>53.83 (32.10)</td>
<td>49.06 (31.60)</td>
<td>t(717)=2.01*</td>
<td>0.15</td>
</tr>
<tr>
<td>Done more, Vehicle B (M)</td>
<td>46.08 (31.78)</td>
<td>40.61 (30.55)</td>
<td>t(717)=2.36*</td>
<td>0.18</td>
</tr>
<tr>
<td>Trust in Autonomous Vehicles</td>
<td>34.35 (23.94)</td>
<td>31.81 (25.25)</td>
<td>t(714)=1.38</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Note: T-statistics reflect results of independent-samples t-tests. * p < .05 ** p < .01, *** p < .001.
Table 4. 
Statistics for Study 3.

<table>
<thead>
<tr>
<th>Measure</th>
<th>AV v HDV Constrained</th>
<th>AV v HDV Unconstrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liability, At Fault (DV)</td>
<td>$t(409)=-0.98$, $d=-0.10$</td>
<td>$t(362)=-2.56^*$, $d=-0.25$</td>
</tr>
<tr>
<td>Firm Liability, Not at Fault (DV)</td>
<td>$t(348)=3.53^{***}$, $d=0.35$</td>
<td>$t(346)=6.61^{***}$, $d=0.66$</td>
</tr>
<tr>
<td>Counterfactual, At Fault (M)</td>
<td>$t(350)=1.25$, $d=0.12$</td>
<td>$t(353)=-0.34$, $d=-0.04$</td>
</tr>
<tr>
<td>Counterfactual, Not at Fault (M)</td>
<td>$t(395)=4.35^{***}$, $d=0.43$</td>
<td>$t(385)=4.52^{***}$, $d=0.46$</td>
</tr>
<tr>
<td>Done more, Not at Fault (M)</td>
<td>$t(394)=3.41^{***}$, $d=0.34$</td>
<td>$t(387)=3.91^{***}$, $d=0.40$</td>
</tr>
</tbody>
</table>

*Note:* T-statistics reflect results of independent-samples t-tests. * $p < .05$ ** $p < .01$, *** $p < .001$.

Table 5. 
Statistics for Study 4.

<table>
<thead>
<tr>
<th>Measure</th>
<th>AV v HDV No Intervention</th>
<th>AV v HDV With Intervention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sue, At Fault (DV)</td>
<td>$t(569)=-2.36^*$, $d=-0.19$</td>
<td>$t(617)=-0.24$, $d=-0.02$</td>
</tr>
<tr>
<td>Sue Firm, Not at Fault (DV)</td>
<td>$t(554)=4.43^{***}$, $d=0.35$</td>
<td>$t(617)=0.45$, $d=0.04$</td>
</tr>
<tr>
<td>Counterfactual, At Fault (M)</td>
<td>$t(617)=-0.65$, $d=-0.05$</td>
<td>$t(548)=0.63$, $d=0.05$</td>
</tr>
<tr>
<td>Counterfactual, Not at Fault (M)</td>
<td>$t(578)=6.16^{***}$, $d=0.49$</td>
<td>$t(618)=2.51^*$, $d=0.20$</td>
</tr>
<tr>
<td>Done more, Not at Fault (M)</td>
<td>$t(600)=5.71^{***}$, $d=0.46$</td>
<td>$t(588)=4.29^{***}$, $d=0.34$</td>
</tr>
</tbody>
</table>

*Note:* T-statistics reflect results of independent-samples t-tests. * $p < .05$ ** $p < .01$, *** $p < .001$. 
**Fig. 1.** Theoretical framework.

![Diagram showing the theoretical framework](image)

**Note:** Proposed thought process for assessing accidents not at fault, by which vehicle type influences whether consumers consider 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 avoided the accident and, ultimately, firm liability for the accident. Individual levels of trust in AVs influence the propensity to engage in these counterfactuals. ‘S’ refers to ‘study’, indicating studies that test each component of the thought process.

**Fig. 2.** Instruction stills for human-driven (left) and autonomous (right) conditions.
Fig. 3. Main results for Study 1.

Fig. 4. Moderated serial mediation for Study 1.

Note: **p < .01; ***p < .001
**Fig. 5.** The effect of trust in AVs on agreement with the counterfactual.

![Agreement Wt. Counterfactual](image)

**Note:** High Trust > 50 average on trust scale; Low Trust ≤ 50. Error bars indicate 95% CIs. ***p < .001.

**Fig. 6.** Scenario and schematics in Study 2.

**Part 1**
Both Vehicle A and Vehicle B wait for the green light. The light turns green, and both vehicles advance shortly thereafter. The pedestrian steps into the crosswalk.

**Part 2**
Vehicle A strikes the pedestrian at a speed of 21.7 mph. Vehicle A appears to hit the pedestrian without applying the brakes.

**Part 3**
The pedestrian lands in Vehicle B’s travel lane.

**Note.** Images and crash details taken from Cano (2024).
Fig. 7. Main results for Study 2.

Fig. 8. The effect of trust in AVs on agreement with the counterfactual, Study 2

Note: High Trust > 50 average on trust scale; Low Trust <= 50. Error bars indicate 95% Cis. * $p < 0.05$ *** $p < .001$. 
Fig. 9. Instruction stills for the agent (human-driven, Top; Autonomous, Bottom) and scenario (unconstrained scenario, left; constrained scenario, right) conditions.

Fig. 10. Instruction stills for agent (human-driven, left column; autonomous, right column) and intervention (no intervention, top row; intervention, bottom row) conditions.