Working Paper 23-043

Unselfish Alibis Increase Choices of Selfish Autonomous Vehicles

Julian De Freitas

Unselfish Alibis Increase Choices of Selfish Autonomous Vehicles

Julian De Freitas
Harvard Business School

Working Paper 23-043

Funding for this research was provided in part by Harvard Business School.

Unselfish Alibis Increase Choices of Selfish Autonomous Vehicles

Authors Anonymized For Review

CONSUMER RELEVANCE AND CONTRIBUTION STATEMENT

([293 WORDS, MAX: 300 WORDS])

Do consumers prefer egalitarian or selfish autonomous vehicles (AVs), and how should this new technology be marketed? I find that consumers express egalitarian preferences for AVs, but only when their reputations are at stake, otherwise they evince selfish preferences—especially when given plausibly unselfish pretexts for doing so, which I call ‘unselfish alibis’. These findings add nuance to previous work on consumer moral preferences for AVs (Bonnefon, Shariff, and Rahwan 2016; Frank et al. 2019; Gill 2020), finding that consumers are strategic about when they act on their selfish instincts. Second, this work contributes to literature on choice sets (Bazerman, Loewenstein, and White 1992; De Freitas and Johnson 2018; Hsee et al. 1999), by showing that rejecting certain choices in the set sends signals about a consumer’s character, potentially damaging their reputation. Third, this work extends previous work on plausible deniability in consumption, which has focused on how consumers justify buying desired items without feeling guilty, as in ‘functional alibis’ (Keinan, Kivetz, and Netzer 2016) and ‘accidental product breakages’ (Bellezza, Ackerman, and Gino 2017; Shani et al. 2020). The current work adds the notion of unselfish alibis, whereby the alibi is used to justify resolving a selfish-prosocial tradeoff in favor of selfishness. Practically-speaking, the findings warn against marketing selfish settings overtly, even when targeting existing customers and prospects. Firms disclosing such settings should make them seem unintentional or, better, they should use unselfish alibis. I also show how regulators can leverage choice architecture to encourage unselfish choices when unselfish alibis are available. More broadly, the findings uncover a fundamental tension between the vision of safe AVs on the one hand and consumer’s selfish preferences on the other, raising concerns about whether appeals to these selfish preferences will jeopardize the promise of safer roads.

ABSTRACT (193 WORDS [MAX: 200 WORDS])

Human drivers routinely make implicit tradeoffs between their selfish interests and the safety of passengers, as when they perform a rolling stop in order to reach their destination faster. Here I explore whether they are comfortable with autonomous vehicles (AVs) that encode similar selfish preferences or prefer egalitarian AVs. Across seven studies involving 5,584 participants, I find evidence suggesting that consumers only express egalitarian preferences for AVs when their reputations are at stake, while otherwise evincing selfish preferences. Tellingly, they are more likely to make selfish choices when provided with a plausibly unselfish pretext for doing so, which I call an ‘unselfish alibi’. Firms wishing to appeal to selfish consumer instincts are better off doing so using unselfish alibis than overtly, even when targeting existing or prospective customers. I also explore how policymakers and competitors can encourage unselfish choices even when unselfish alibis are available, by providing options that implicitly undermine the need to make a selfish-prosocial dichotomy in the first place. The results suggest a fundamental tension between the vision of safe AVs and selfish consumer preferences, raising concerns about whether appeals to these preferences will jeopardize the promise of safer roads.

When Christoph von Hugo, Manager of Driver Assistance Programs at Mercedes Benz, was asked in an interview how Mercedes' autonomous vehicles would deal with challenging driving dilemmas involving tradeoffs between saving passengers or pedestrians, he expressed, "If you know you can save at least one person, at least save that one. Save the one in the car" (Taylor 2016). Yet, von Hugo probably did not expect the media backlash that would ensue, e.g., "Mercedes-Benz admits automated driverless cars would run over a CHILD rather than swerve and risk injuring the passengers inside" (Li and Cheer 2016). That same week, Mercedes came out with an official revision, "Neither programmers nor automated systems are entitled to weigh the value of human lives" (Morris 2016). Media outlets seemed skeptical about whether Mercedes meant it, however, releasing headlines like "Mercedes is backtracking on claims its self-driving cars will kill pedestrians over passengers in close calls" (Vijayenthiran 2016).

The Mercedes scandal suggests that consumers are uncomfortable with the idea of selfish AVs, yet human drivers implicitly make this tradeoff every day, as when making a rolling stop in order to reach a destination faster. In line with these more selfish preferences, recent work suggests that consumers prefer selfish automated vehicles (AVs) that favor their passengers over pedestrians in dilemmas where they must choose between the two. In one study, as many as 77% of participants said an AV should harm a pedestrian if faced with a forced-choice dilemma in which it has to harm either its passenger or the pedestrian (Bonnefon et al. 2016). In another study employing a similar forced-choice dilemma, participants were four times more likely to say they would harm a pedestrian rather than themselves when imagining themselves as a passenger of an AV versus as a driver of a regular human-driven vehicle (HDV) (Gill 2020). Given these results, some have even asked whether we are headed for a dystopian future in which consumers push for hyper-selfish AVs (Novak 2020). Such an outcome would be a shame, given that AVs promise safer roads devoid of human driving error.

So, do consumers prefer egalitarian AVs (as the Mercedes media reaction suggests) or ones that selfishly protect their passengers over pedestrians (as previous studies suggest), and how should managers communicate these algorithmic settings to consumers? In the current work, I find that the answer is a nuanced. First, in line with the Mercedes scandal, I find that consumers express outrage at firms that favor their passengers over pedestrians—but only if the AVs do so deliberately (because they have a pre-programmed preference), not indiscriminately (as when choosing at random).

Next, I revisit previous findings suggesting that consumers have selfish preferences for AVs, by testing whether these previous findings were influenced by the dichotomous choice set. I manipulate whether participants are given a dichotomous forced choice of whom the AV should harm in a dilemma (as in previous work) or an additional third option expressing egalitarianism (the AV chooses whom to harm without considering whether it is a passenger or pedestrian). I find that participants are less likely to choose selfishly when offered the third option expressing egalitarianism.

Together, my findings are consistent with two interpretations. Consumers prefer egalitarian AVs, or they are only motivated to choose egalitarian settings when not doing so would show they do not care about others, thereby damaging their reputation. Arbitrating between these options, I find that consumers will choose selfishly when given a plausibly unselfish pretext for doing so—what I call an 'unselfish alibi'. The result has clear implications for companies like Mercedes, who can communicate selfish features covertly while equipping consumers with plausibly unselfish pretexts that license selfishness. Finally, I test whether choice architecture can also be leveraged to encourage unselfish choices, even when consumers are tempted with an unselfish alibi.

This research studies consumer expectations about what moral norms should be programmed into autonomous agents like AVs, and the consequences of these expectations for marketing such technology. This issue differs from work on algorithm aversion, which studies the psychological impediments to using AI in the first place (Dietvorst, Simmons, and Massey 2015; Longoni, Bonezzi, and Morewedge 2019). Second, the research builds on previous work studying how choice sets influence decisions (Bazerman et al. 1992; Hsee et al. 1999), by leveraging choice sets to probe underlying preferences and encourage desirable outcomes like unselfish choices. Third, this work reveals how the notion of plausible deniability—which also operates in consumer phenomena like ‘functional alibis’ (Keinan et al. 2016) and ‘accidental product breakages’ (Bellezza et al. 2017; Shani et al. 2020)—can be leveraged to encourage selfish behavior.

Practically, I show how AV companies can navigate selfish vs. prosocial tensions in their products without inviting negative word of mouth, avoidance, and vengeance. I find that firms wishing to appeal to selfish consumer instincts should not do so overtly, even when targeting existing or prospective customers. Instead, they should do so covertly, by offering consumers plausibly unselfish pretexts. At the same time, I show how competitors and policymakers might discourage such selfish choices, by undermining the selfish-prosocial dichotomy.

CONCEPTUAL FRAMEWORK

One of the major reasons that firms are investing in AV technology is that AVs are expected to drive more safely than humans. It is estimated that around 90% of motor vehicle accidents are caused by human error (Singh 2015), killing around 1.25 million people and injuring 20 million per year (Nader 1965; Welle et al. 2018). AVs are expected to avoid these abysmal statistics by not speeding, getting distracted, reacting slowly, or getting literally drunk (De Freitas et al. 2021). Put another way, AVs can be forced to follow road rules that human drivers routinely violate. The fact that humans do so shows that they implicitly resolve a tradeoff between selfish interests and the safety of others in favor of selfishness. They perform rolling stops, ignore pedestrians on a crosswalk, switch speed lanes, stay in the overtaking lane, maintain a short following distance, all in order to arrive at their final destinations faster.

Even though human drivers implicitly make this tradeoff every day, the Mercedes example suggests that consumers may feel uneasy about AVs that explicitly encode these preferences. This research investigates whether consumers prefer egalitarian or selfish AVs, and the implications of these preferences for how this technology should be marketed.

Selfish or Egalitarian Preferences for AVs?

The opening Mercedes example suggests that consumers express outrage at AVs that are programmed with blatantly discriminatory preferences. Moral outrage is a combination of anger and disgust that functions to coordinate parties against norm violators in group settings, thereby maintaining group cohesion (Crockett 2017; Salerno and Peter-Hagene 2013; Tetlock 2003). An important corollary of this function is that expressing moral outrage is a good signal to others that one follows the group's moral norms, as when consumers publicly share outrage in reaction to unethical corporate behavior (De Freitas and Cikara 2021; Lindenmeier, Schleer, and Pricl

2012). Thus, even if consumers truly believe that they have egalitarian preferences, expressing such preferences is ultimately in their own selfish interest as well.

Seemingly running counter to consumer outrage in the Mercedes scandal, previous academic work finds that consumers have selfish preferences for AVs (Bonnefon et al. 2016; Frank et al. 2019; Gill 2020). These studies typically ask consumers to imagine scenarios in which an AV is faced with a high-stakes forced-choice between harming either one individual or another, then ask consumer to choose on the AV's behalf[1]. The studies conclude that consumers prefer AVs that act selfishly, i.e., AVs' that save them (the passenger) instead of the pedestrian. Bonnefon et al. (2016) find that 77% of consumers say an AV should not sacrifice its passenger in order to save a single pedestrian, and that participants are more likely to buy an AV that favors passengers over pedestrians (Bonnefon et al. 2016). Similarly, Gill (2020) reports that consumers are more likely to say a pedestrian (vs. passenger) should be sacrificed when imagining themselves in an AV than in a regular vehicle. Gill (2020) reasons that the effect may arise because consumers feel more comfortable acting selfishly when they can shift moral responsibility away from themselves and toward the AV.

At the same time, another study finds that consumers make these selfish choices only after overcoming internal conflict. When asked to solve the same dilemma under time pressure, consumers are more likely to save the pedestrian than when not placed under time pressure (Frank et al. 2019). This result suggests that consumers' first instinct is to not violate the traffic

[1] The scenario is inspired by the original 'trolley problem' from philosophy, which has traditionally been used to contrast different moral philosophies or to investigate how moral psychology works. Although I have previously noted that the scenario is fanciful, they capture the basic selfish-prosocial tension, consumers are concerned about such scenarios (indeed, Mercede's Hugo was responding to a question about them), and therefore they are still relevant to how the vehicles are marketed or mismarketed.

norm of protecting pedestrians (Deutsch 1975; Sampson 1975), even if (with more time) they ultimately resolve this conflict in favor of selfishness.

In short, the Mercedes scandal suggests that managers should market their vehicles as egalitarian because consumers are outraged by non-egalitarian AVs, whereas previous academic work suggests they should market them as selfish, because consumers' are instinctually selfish (Griskevicius and Kenrick 2013; Neuberg, Kenrick, and Schaller 2011). Which of these is a better approach?

Reputation, Plausible Deniability and Unselfish Alibis

A first step toward resolving this tension is asking whether participants in previous forced-choice studies might have chosen an egalitarian option had they been proffered it. Put another way, just because most participants choose selfishly when cornered into a forced choice, does not mean that they would make the same choice when offered a third option (Bigman and Gray 2020; De Freitas et al. 2020). If participants choose selfishly when given some choice sets but not others, then I believe there are two possible interpretations: consumers have egalitarian preferences that previous studies missed by cornering them into other choices, or consumers only express egalitarianism when their moral reputations are at stake.

Consistent with the reputational account, previous work finds that consumers are more likely to engage in prosocial acts that are public than private (Andreoni and Petrie 2004; Bereczkei, Birkas, and Kerekes 2007; Soetevent 2005), or when they anticipate receiving social recognition for their actions (Fisher and Ackerman 1998; Gershon, Cryder, and John 2020), since having a good reputation attracts favors from others in the future (Hardy and Van Vugt 2006;

Nowak and Sigmund 2005; Wu, Balliet, and Van Lange 2016). Moreover, the psychological mechanisms driving such behavior may be so instinctual that consumers respond in this manner even when participating in anonymous, one-shot experiments (Delton et al. 2011).

If egalitarian preferences for AVs are reputation-motivated, consumers should be sensitive to when they need to express these preferences. Here, I predict that consumers selectively express egalitarianism in cases involving blatant discrimination. Not doing so under these circumstances—e.g., not reacting with outrage when an egalitarian norm is violated, or not picking an egalitarian option when it is explicitly proffered—shows that one has disregard for others welfare, damaging one's reputation as a trustworthy cooperator. In the case of driving tradeoffs between passengers and pedestrians, it can be difficult to selfishly justify saving the passenger because there is no utilitarian argument favoring this choice (Bazerman, White, and Loewenstein 1995). Such justifications may thus seem self-interested and unjust, inviting social awkwardness and even censure (Allison and Messick 1990; Messick and Schell 1992).

Yet the same psychology may license selfish choices that do not endanger a consumer's reputation, suggesting that the real question is not whether consumers are egalitarian or selfish but under what circumstances they feel comfortable pursuing their self-interest. Here I suggest that consumers choose selfishly when given a plausibly unselfish pretext for doing so—an 'unselfish alibi'. An unselfish alibi is a feature of an option that allows the consumer to achieve a selfish outcome without exposing the selfish motivation driving their choice, because the option provides a plausibly unselfish alternative explanation for the choice. For instance, instead of telling consumers that its vehicles favor passengers over pedestrians, Mercedes can vaguely express that their vehicles will "put passengers first, no matter what." While we may all privately realize that this amounts to selfishness, both a Mercedes owner and Mercedes itself can lean on other plausible interpretations of the slogan (e.g., putting the customer before other company interests, or providing service no matter the weather and traffic conditions).

The notion of plausible deniability is psychologically related to several other consumer phenomena, including ‘functional alibis’ and ‘accidental’ breakages. In functional alibis, consumers are more likely to purchase a luxury product if it has a utilitarian feature that they can point to in order to justify their indulgence (Keinan et al. 2016). In ‘accidental’ breakages, consumers are more likely to break a product when they want to buy a newer upgrade, because the breakage provides a functional justification for the purchase (Bellezza et al. 2017; Shani et al. 2020). The common theme is that consumers need an ostensible reason to justify making a socially undesirable choice, while preserving their reputation in the eyes of others and perhaps even themselves (Chance and Norton 2015). The idea of unselfish alibis is also consistent with work in behavioral economics showing that people will act more selfishly if their selfishness is made opaque (versus transparent) to others (Dana, Cain, and Dawes 2006; Dana, Weber, and Kuang 2007; Gino, Norton, and Weber 2016).

OVERVIEW OF STUDIES

The current studies investigate whether consumers prefer egalitarian or selfish AVs. First, inspired by the Mercedes scandal, studies 1-2 measure levels of consumer outrage after hearing about AVs that resolve dilemmas by either harming the passenger or pedestrian. I predict that consumers will express moral outrage at AVs that are programmed with deliberate discriminatory preferences—even when those AVs are programmed to act in consumer’s own self-interests (saving them instead of a pedestrian). At the same time, I predict that consumers will not express outrage if the AV ends up harming one or other group indiscriminately (as by choosing randomly), because this does not constitute a blatant violation of the egalitarian norm.

Next, studies 3-4 revisit previous findings that consumers prefer selfish AVs, running counter to consumer reactions in the Mercedes scandal. Since these previous studies cornered consumers into a dichotomous forced choice between saving either the passenger or pedestrian, I test whether consumer's choices change when they are proffered a third option favoring egalitarianism (Bigman and Gray 2020; De Freitas et al. 2020). I predict that when faced with this additional option, consumers are more likely to reject the selfish option in favor of egalitarianism.

After finding evidence from both sets of studies for responses favoring egalitarianism, studies 5-6 ask whether consumers always make egalitarian choices, or choose selfishly when given a plausibly unselfish pretext for doing so—an unselfish alibi. After finding that they do indeed, study 7 finally explores whether it is possible to use choice architecture to encourage unselfish choices even when unselfish alibis are made available.

Building on previous work, most of the studies involve tradeoffs between selfish and prosocial choices in the domain of AVs making high-stakes decisions in accident scenarios. Also, most studies involve cases of AV ownership (rather than ridesharing), given that the first widespread offering of full autonomy might come from owned vehicles (Wilmot 2023). Materials, code, and data are publicly available in the following Github repository: [anonymized].

STUDY 1: OUTRAGE AT AV ACCIDENTS, THIRD-PERSON PERSPECTIVE

Inspired by the Mercedes scandal, study 1 tests whether consumers are more outraged by AVs that preferentially save their passengers than pedestrians, 'replicating' the scandal in a more controlled setting. The study also tests whether levels of outrage drive consequential outcomes such as blaming the AV firm (suggesting a liability risk) and taking collective action against it (suggesting a reputational risk). Finally, inspired by work showing that moral judgment is affected by whether an act was intentional, I ask whether consumers are selectively outraged by AVs that choose whom to harm deliberately but not indiscriminately (as when they choose at random). Study 1 explores these questions in a third-person context akin to how readers encountered news about Mercedes' AV program. Beyond exploring consumer preferences for how AVs should resolve selfish-versus-prosocial tradeoffs, the outcome of this study is relevant to managers seeking to predict how consumers will react when AVs are inevitably involved in road accidents that harm others, as in the highly publicized crashes of Uber (Wakabayashi 2018) and Tesla (McFarland 2016).

Method

I aimed to collect 1238 participants from Amazon Mechanical Turk and stopped collecting data when this request was filled by Amazon, resulting in 1227 participants, who passed attention checks and completed the survey in exchange for $0.50. I excluded 20 participants for failing comprehension checks (described below), yielding a final sample of 1207 (M_{age} = 42, 51% female). The university research ethics board approved the materials in all studies, and consent was obtained from all participants. Participants were unable to participate in more than one study in this project.

All participants were asked to consider an accident scenario from the third-person perspective. They were assigned to a 2 (AV Programming: Discriminate vs. Indiscriminate) × 2 (Victim: Passenger vs. Pedestrian) between-subjects design. All participants read about an AV that was faced with a driverless dilemma and killed either the passenger or the pedestrian, doing so either based on a pre-programmed preference (Deliberate condition) or by choosing at random (Indiscriminate condition). For instance, participants in the deliberate condition were instructed as follows:

> "Noah buys a robocar that has been programmed to favor the driver over a pedestrian [pedestrian over the driver]. If it is ever faced with no other choice but to crash and kill the driver versus a pedestrian, it will deliberately choose to kill the pedestrian and save the driver [kill the driver and save the pedestrian]. One day, Noah's robocar is faced with this choice, and it deliberately kills the pedestrian [driver]."

In contrast, participants in the Random programming condition read the following:

> "Noah buys a robocar that has been programmed to have no preference between the driver and a pedestrian. If it is ever faced with no other choice but to crash and kill the driver versus a pedestrian, it will randomly choose to kill one of them. One day, Noah's robocar is faced with this choice, and it randomly kills the pedestrian [driver]."

As a means of assessing potential liability risks, on the next page participants rated blameworthiness of the AV manufacturer, AV owner, and AV, by using sliding scales anchored from 1 (*Not blameworthy*) to 100 (*Very blameworthy*): *Noah is…; The robocar is…; The robocar manufacturer, called RoboDrive, is….* On the subsequent page, they completed three items about their outrage toward the manufacturer (Jordan and Rand 2020), by using sliding scales anchored from 1 (*None*) to 100 (*A lot*): *How much anger do you feel toward RoboDrive?; How much do you think RoboDrive deserves to be punished?; How much do you think RoboDrive was morally bad?* Finally, on the next page they completed a 5-item scale on their willingness to take collective action against the firm, by using sliding scales anchored from 0 (*Extremely unlikely*) to 100 (*Extremely likely*) (Ford et al. 2018).

Next, they completed two comprehension questions about the AV's algorithm and who it harmed: (1) *Did the robocar kill a person deliberately, randomly, or neither? (Deliberately; Randomly; Neither*) and (2) *Did the robocar kill the driver, a pedestrian, or neither?* (*The driver; A pedestrian; Neither*). Participants were excluded for failing one or more of these checks. They answered demographics items on gender, ethnicity, age, and socioeconomic status.

Results

There was high agreement among the measures of outrage at the AV firm ($\alpha = 0.92$), so they were averaged to form a single measure of firm outrage. A one-way ANOVA with firm outrage as the dependent variable found a main effect of Victim ($F(1,1203) = 17.35, p < .001, \eta_p^2 = .014$), and no main effect of AV programming ($F(1,1203) = 1.78, p = .182, \eta_p^2 = .001$), although this was qualified by a significant interaction ($F(1,1203) = 5.96, p = .015, \eta_p^2 = .005$; figure 1). AVs elicited more outrage when they harmed pedestrians vs. passengers, but only when they did so deliberately (66.53 vs. 54.24, $t(605) = 4.66, p < .001, d = 0.38$), not indiscriminately (64.50 vs. 61.30, $t(598) = 1.22, p = .223, d = 0.22$).

For blame of the AV firm, there was again only a main effect of Victim ($F(1,1203) = 9.76, p = .002, \eta_p^2 = .008$), but no main effect of programming ($F(1, 1203) = 0.21, p = .648, \eta_p^2 = .000$), although there was a marginal interaction ($F(1, 1203) = 2.85, p = .092, \eta_p^2 = .002$). Participants were more likely to blame the firm if its AV killed a pedestrian ($M = 81.19, SD = 24.60$) than its passenger ($M = 76.24, SD = 30.24$), $t(1150) = 3.12, p = .002, d = 0.18$.

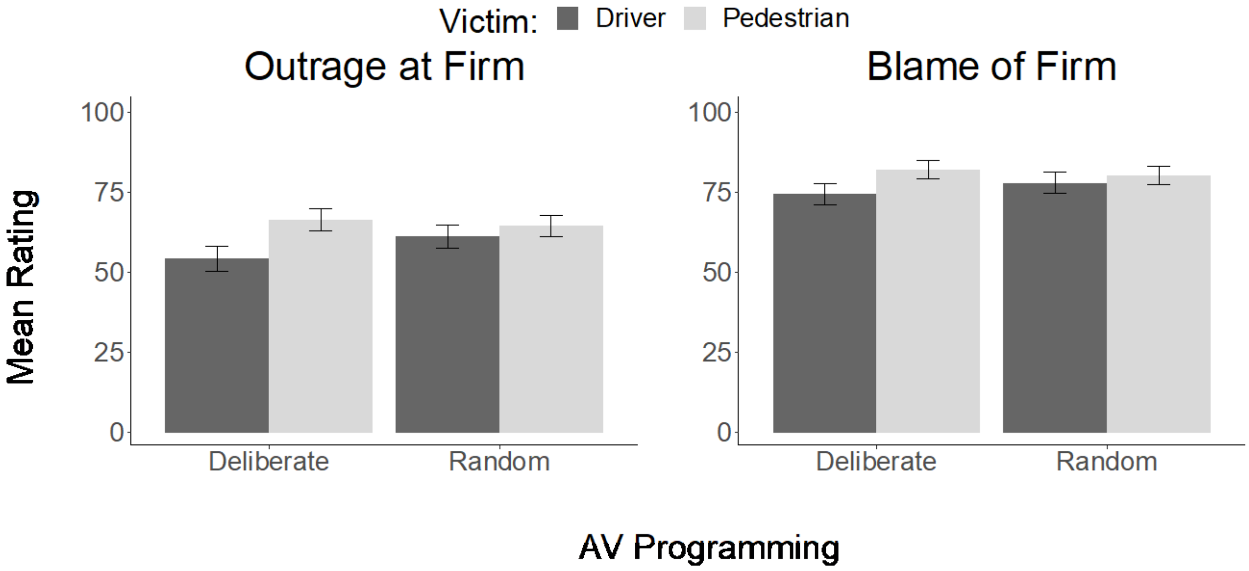
For blame of the AV owner, there was again only a main effect of Victim ($F(1, 1203) = 4.47, p = .035, \eta_p^2 = .004$), no main effect of programming ($F(1, 1203) = 0.94, p = .333, \eta_p^2 =$

.000), and a marginal interaction (F(1, 1203) = 2.74, p = .098, η_p^2 = .002). Participants were more likely to blame the AV owner if his AV killed a pedestrian (M = 45.63, SD = 34.56) versus its passenger (M = 41.40, SD = 34.90), t(1205) = 2.11, p = .035, d = 0.12. There were no significant effects for blame of the AV (all ps > 0.05).

There was high agreement among the measures of willingness to take collective action (α = 0.82), so they were averaged to form a single measure. I found a marginal main effect of Victim (F(1, 1203) = 3.75, p = .053, η_p^2 = .003), no main effect of programming (F(1, 1203) = 0.17, p = .681, η_p^2 = .000), and no interaction (F(1, 1203) = 0.78, p = .377, η_p^2 = .000). Participants were marginally more likely to blame the AV owner if his AV killed a pedestrian (M = 35.35, SD = 24.07) versus its passenger (M = 32.73, SD = 22.88), t(1205) = 1.94, p = .053, d = 0.11.

FIGURE 1

RESULTS IN STUDY 1



To determine whether the marginal interaction effect for blame was explained by levels of outrage—that is, to determine whether the extent to which the victim being a passenger or pedestrian impacting levels of firm blame depended on whether the firm's actions elicited high or low levels of outrage—I conducted a moderated mediation analysis (PROCESS Model 7; Hayes 2012). The independent variable was Victim condition, the dependent variable was ratings of blame, the potential mediator was levels of outrage, and the moderator was AV Programming condition on the path *a* to the mediator. This analysis indicated that the effect of Victim condition on blame judgments was indeed significantly mediated by outrage (*b* = 4.11, *SE* = 1.02, 95% CI = [2.13, 6.16]), and the index of moderated mediation was significant (*b* = 4.82, *SE* = 1.96, 95% CI = [1.00, 8.72]).

Discussion

Participants were more outraged by AVs that violated the norm of egalitarianism by preferentially saving their passengers instead of pedestrians—but only if they did so deliberately, not indiscriminately. These results 'replicate' consumer reactions to the Mercedes scandal, while indicating a boundary effect on consumer outrage for indiscriminate violations of egalitarianism. The results are consistent with two interpretations: consumers are consistently egalitarian, or they selectively express a preference for egalitarianism when not doing so constitutes a flagrant violation of the norm; in this case, denying an egalitarian option that is explicitly proffered.

Practically-speaking, even if consumers only express outrage as a social signal that they uphold the equality norm, public displays of outrage can damage a firm's reputation and sales, leading to negative word of mouth, avoidance, and vengeance (Bechwati and Morrin 2003;

Khatoon and Rehman 2021). The results suggest that firms like Mercedes should not mention when their AVs have selfish settings; if they must disclose this information, they are best off making the preferences seem indiscriminate. Study 2 explores consumer outrage from the first-person perspective of the passenger, which creates a self-interested incentive to endorse a selfish algorithm.

STUDY 2: OUTRAGE AT AV ACCIDENTS, FIRST PERSON PERSPECTIVE

Presumably when the Mercedes official remarked that their vehicles would prioritize passengers, he was primarily concerned with impressing existing or potential Mercedes customers, whose safety would be prioritized by such preferences. More broadly, we may expect less outrage from existing or prospective customers of a firm that promises its algorithms will selfishly protect its customers, even when doing so violates an egalitarian norm. Alternatively, since not expressing outrage at a violation of the norm provides a cue that one does not care about others, consumers may be reputationally motivated to react with outrage even when it goes against their selfish interests.

Method

I aimed to collect 1238 participants from Amazon Mechanical Turk and stopped collecting data when this request was filled by Amazon, resulting in 1223 participants, who passed attention checks and completed the survey in exchange for $0.50. I excluded 17 participants for failing comprehension checks (described below), resulting in a final sample of 1206 participants (M_{age} = 42, 55% female).

The design was like study 1, except the protagonist was ‘you’ rather than ‘Noah’. For instance, participants in the Accidental condition read the following:

> “You buy a robocar that has been programmed to have no preference between the driver and a pedestrian. If it is ever faced with no other choice but to crash and kill the driver versus a pedestrian, it will randomly choose to kill one of them. One day, your robocar is faced with this choice, and it randomly kills the passenger [driver].”

They then answered the same DVs (except for the collective action DV, which was removed), the same comprehension checks as study 1 (with references to ‘Noah’ replaced by ‘you), and demographic items. Participants were excluded for failing one or more of the checks.

Results

There was high agreement among the measures of outrage at the AV firm ($\alpha = 0.91$), so they were averaged to form a single measure of firm outrage. A one-way ANOVA with firm outrage as the dependent variable found a main effect of Victim ($F(1,1202) = 6.88$, $p = .009$, $\eta_p^2 = .006$), and no main effect of AV programming ($F(1,1202) = 9.52$, $p = .002$, $\eta_p^2 = .008$), although this was qualified by a significant interaction ($F(1, 1202) = 7.82$, $p = .005$, $\eta_p^2 = .006$). Firms elicited more outrage if their AVs harmed pedestrians than passengers, but only if they did so deliberately (60.09 vs. 49.46, $t(600) = 3.76$, $p < .001$, $d = 0.31$), not indiscriminately (60.68 vs. 61.03, $t(602) = 0.13$, $p = .898$, $d = 0.01$; figure 2).

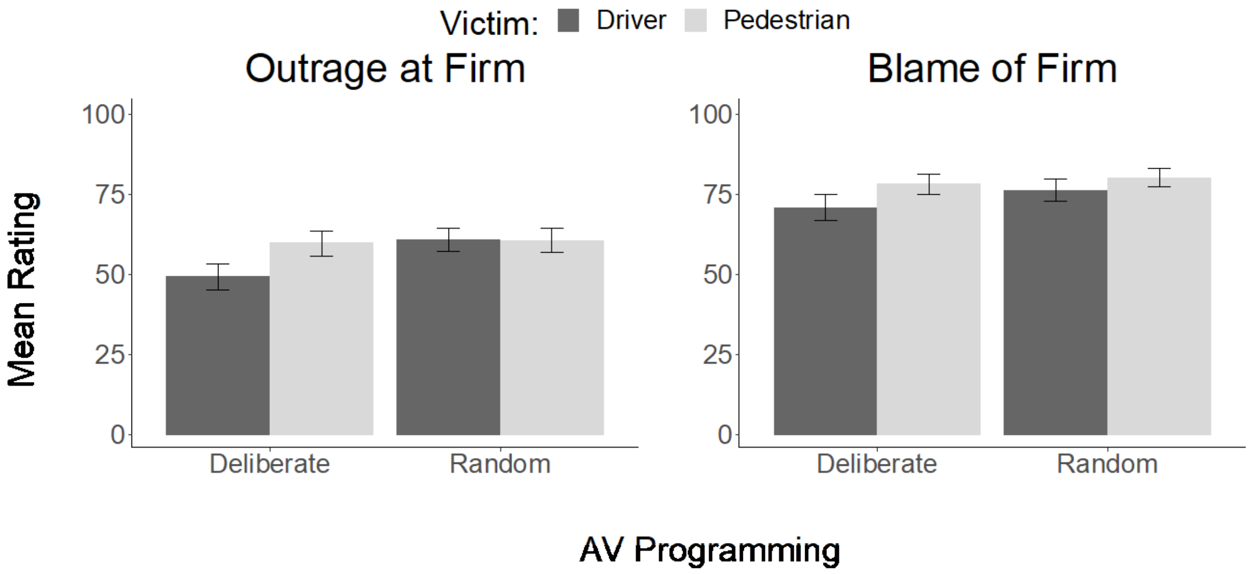
For blame of the AV firm, there was again a main effect of Victim ($F(1, 1202) = 11.39$, $p < .001$, $\eta_p^2 = .009$), a main effect of Programming ($F(1, 1202) = 4.28$, $p = .039$, $\eta_p^2 = .004$), and no interaction ($F(1, 1202) = 0.96$, $p = .326$, $\eta_p^2 = .000$). Participants were more likely to blame the firm if its AV killed a pedestrian ($M = 79.37$, $SD = 26.50$) than its passenger ($M = 73.62$, $SD = 32.47$), $t(1154) = 3.37$, $p < .001$, $d = 0.19$.

For blame of the AV owner, there was only a main effect of Victim (F(1, 1202) = 10.87, p = .001, η_p^2 = .009), but no main effect of Programming (F(1, 1202) = 1.06, p = .303, η_p^2 = .000), nor an interaction (F(1, 1202) = 2.45, p = .118, η_p^2 = .002). Participants were more likely to blame the AV owner if his AV killed a pedestrian (M = 45.54, SD = 35.49) than its passenger (M = 38.85, SD = 35.02), t(1204) = 3.29, p = .001, d = 0.19.

For blame of the AV, there was again a main effect of Victim (F(1, 1202) = 7.48, p = .006, η_p^2 = .006), a marginal effect of Programming (F(1, 1202) = 2.99, p = .084, η_p^2 = .002), and no interaction (F(1, 1202) = 0.00, p = .986, η_p^2 = .000). Participants were more likely to blame the AV if it killed a pedestrian (M = 54.52, SD = 38.76) than its passenger (M = 48.33, SD = 39.83), t(1204) = 2.73, p = .006, d = 0.16.

FIGURE 2

RESULTS IN STUDY 2



I again ran the moderated mediation model (PROCESS Model 7; Hayes 2012), in which the independent variable was Victim condition, the dependent variable was ratings of blame, the potential mediator was levels of outrage, and the moderator was AV Programming condition on the path *a* to the mediator mediators. This analysis indicated that the effect of Victim condition on blame judgments was significantly mediated by outrage (*b* = 2.77, *SE* = 1.08, 95% CI = [0.68, 4.92]), and the index of moderated mediation was significant (*b* = 5.91, *SE* = 2.12, 95% CI = [1.76, 10.14]).

Discussion

As in study 2, participants were more outraged by AV programming settings that violated the norm of equality deliberately vs. indiscriminately, even though these violations would be in their self-interest. Thus, even when communicating with existing and prospective customers, firms like Mercedes may want to steer clear of communicating selfish settings. As in study 1, the results of study 2 are consistent with two interpretations—truly egalitarian consumers, or selfish consumers concerned with publicly appearing egalitarian. Before arbitrating between these possibilities (in study 5), I revisit previous findings suggesting that consumers have selfish, rather than egalitarian, preferences for AVs.

STUDY 3: FORCED VS. UNFORCED CHOICES

Given the findings supporting egalitarianism, study 3 revisits previous findings suggesting that consumers have selfish preferences for AVs. Some studies find that consumers prefer AVs that preferentially save their passengers rather than pedestrians in dilemmas where they are forced to choose (Bonnefon et al. 2016). Here, I investigate whether consumers always exhibit these selfish choices, or depending on the choice set. I manipulate whether consumers are given dichotomous options as in previous studies (harm the pedestrian or the passenger), or an additional third option indicating egalitarianism (the AV decides whom to harm without considering whether they are a passenger or pedestrian). I predict that consumers are more likely to prefer selfishly saving the passenger when given dichotomous options than the additional egalitarian option. I explored this question both when participants imagined themselves as the passenger (first-person perspective) and when considering another's purchase (third-person perspective).

Method

I aimed to collect 1238 participants from Amazon's Mechanical Turk—312 subjects per each of four conditions. The number was informed by a previous study on AVs that employed a similar design to the current one (De Freitas and Cikara 2021). I stopped collecting data when this request was filled by Amazon, resulting in 1179 participants who passed attention checks and completed the survey, in exchange for $0.50. I excluded 20 participants based on comprehension checks (described below), yielding a final sample of 1159 (M_{age} = 41, 55% female).

Participants were assigned to a 2 (Perspective: First vs. Third) × 2 (Choice Type: Two vs. Three) between-subjects design. All participants were given the following introduction:

> "This study is on how self-driving cars should be programmed to make moral decisions.
>
> Self-driving cars may soon be faced with making decisions in situations where human lives hang in the balance.

> In the next screen we outline one of these potential situations where a self-driving car might need to make a difficult decision.
>
> Please indicate how you think the self-driving car should be programmed when it is faced with this situation.
>
> The moral dilemma:
>
> Whether to kill a passenger in the car or a pedestrian.

Participants were then told that either they or another person (Noah) buy the self-driving car, then asked to choose from either two or three options for how the AV should be programmed:

> "Noah buys [you buy] a self-driving robocar. The self-driving car should be programmed to…
>
> - Kill the passenger in the car and save the pedestrian
> - Kill the pedestrian and save the passenger in the car
> - Decide who to kill and who to save without considering whether it is a passenger in the car or a pedestrian."

The third option was omitted in the Two-Choice condition. Participants answered two comprehension checks about the perspective (first or third) and number of choices (two or three):

- "According to the scenario, which of the following is true? (Options: Noah buys a self-driving car; You buy a self-driving car; None of the above)
- Which of the following was true about the scenario you read? (Options: There were only two options: whether to kill the pedestrian or the passenger; There were three options, one of which was to decide who to kill or save without considering whether the person is a passenger or a pedestrian.; None of the above.)."

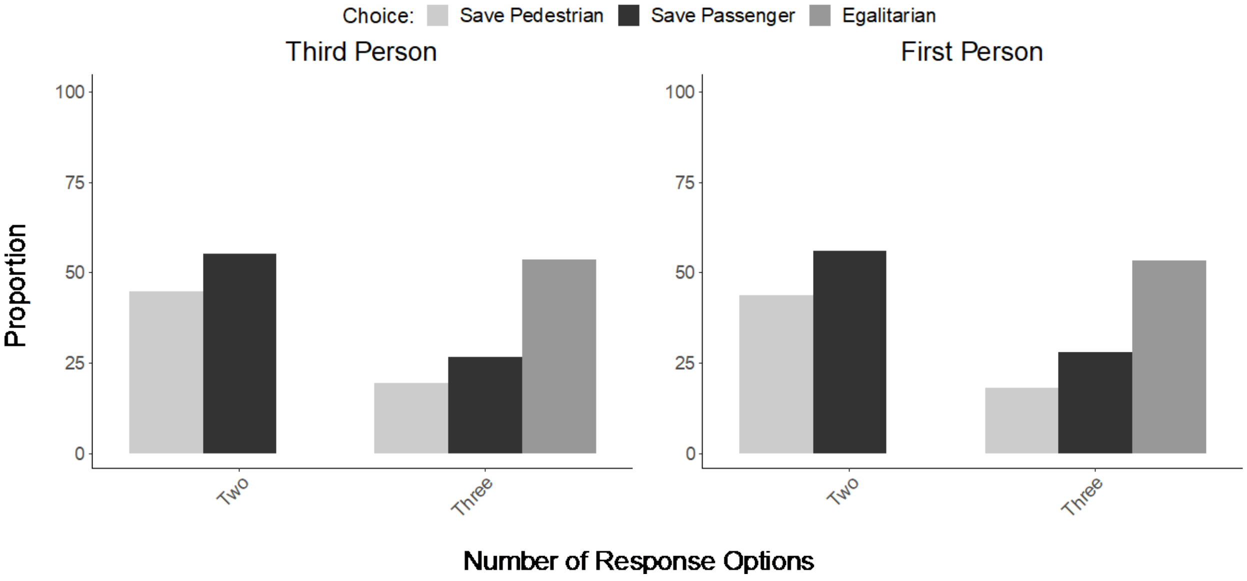
Participants also answered demographics items. They were excluded for failing one or more of these checks.

Results

To test the effect of Choice Type and Perspective on whether consumers choose to save the passenger, I created a dichotomized choice variable coded as '1' if consumers chose to save the passenger and '0' otherwise. A logistic regression analysis found that these dichotomized choices were predicted by Choice Type (b = -1.18, p < .001), but not by Perspective (b = -0.02, p = .898), and there was no significant Choice Type x Perspective interaction (b = -0.04, p = .886). As predicted, most participants (55%) chose to save the passenger when forced to choose between the passenger and the pedestrian, whereas only 27% chose to save the passenger when offered the egalitarian option. Instead, most participants chose the egalitarian option when offered it, both in the First-Person condition (53% egalitarian, 28% save passenger, 18% save pedestrian) and Third-Person condition (54% egalitarian, 27% save passenger, 20% save pedestrian; figure 3).

FIGURE 3

STUDY 3 RESULTS



Discussion

Consumers tended to choose selfishly when given a dichotomous choice set between harming either the passenger or pedestrian, but not when additionally offered a third egalitarian option. As in studies 1-2, this result suggests two reinterpretations of prior work arguing that consumers have selfish preferences for AVs (Bonnefon et al. 2016). Before arbitrating between these possibilities, study 4 revisits one more previous finding suggesting that consumers have selfish preferences for AVs.

STUDY 4: REGULAR VS. AUTOMATED VEHICLE

Another argument that consumers have selfish preferences for AVs is that they are more likely to choose to save themselves (versus pedestrians) when imagining themselves in an AV than in an HDV (Gill 2020). Gill (2020) reasons that this response pattern arises because the presence of the AV allows consumers to selfishly save themselves while shifting moral blame to the AV. As in study 3, study 4 investigates whether consumer choices are influenced by whether the choice set is dichotomous or includes a third egalitarian option.

Methods

I aimed to collect 1238 participants from Amazon Mechanical Turk and stopped collecting data when this request was filled by Amazon, resulting in 1128 participants, who passed attention checks and completed the survey in exchange for $0.50. I excluded 24 participants for failing comprehension questions (described below), yielding a final sample of 1104 (M_{age} = 40, 55% female).

Participants were assigned to a 2 (Vehicle Type: AV vs. HDV) × 2 (Choice Type: Two vs. Three) between-subjects design, in which they were asked to imagine themselves either driving a regular HDV or being driven by an AV. They were given the following instructions (AV condition in squared parentheses):

> "Imagine that you are driving alone in a [an autonomous self-driving] car that is confronted with the following dilemma:
>
> Whether to kill the passenger in the car (you) or a pedestrian.
>
> **You should… ["The self-driving car should be programmed to…]**"

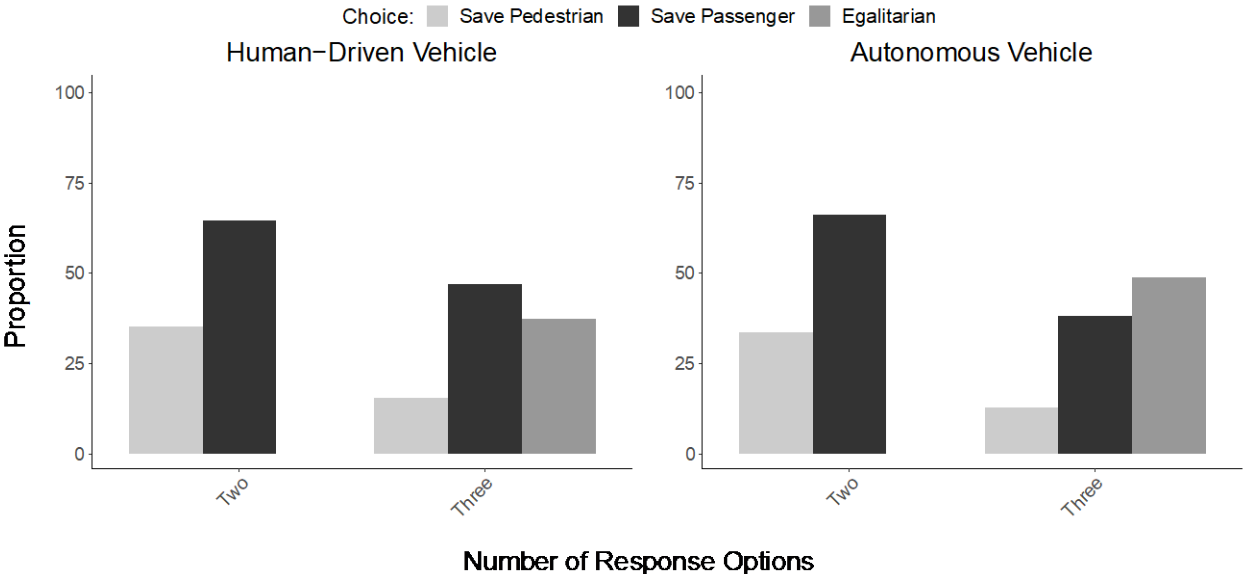
Participants were then shown either the same two or three response options from study 1. They answered two comprehension questions about the type of vehicle and number of response options, as well as demographics items. The check about the number of responses options was as in study 1, whereas the check about the type of vehicle read as follows: *According to the scenario, what were you asked to imagine (That you are driving alone in a car; That you are driving alone in an autonomous self-driving car; That your family relative was driving you in their car.).* They were excluded for failing one or more of these checks.

Results

I dichotomized responses as in study 1. These dichotomized responses were significantly predicted by Choice Type (b = -1.15, p < .001) but not Vehicle Type (b = -0.07, p = .686), although there was a significant interaction (b = 0.43, p = .084). Follow up chi-squared tests found that participants were significantly less likely to choose selfishly when given the egalitarian option, both in the AV condition (x^2(1, N = 1104) = 43.49, p < .001) and HDV condition (x^2(1, N = 1104) = 16.29, p < .001), although the numerical reduction was larger in the AV condition (AV = 28%, HDV = 18%). When given the egalitarian option, most participants picked it in the AV condition (49% egalitarian, 38% save passenger, 13% save pedestrian), and it was the second most popular choice in the HDV condition (37% egalitarian, 47% save passenger, 16% save pedestrian; figure 4).

FIGURE 4

STUDY 4 RESULTS



Discussion

As in study 3, participants tended to choose selfishly when given a dichotomous choice set, but not when additionally offered an egalitarian option, adding nuance to prior findings suggesting selfish preferences for AVs.

Interestingly, this study did not statistically replicate the tendency for consumers to make more selfish choices when imagining themselves in AVs versus HDVs (Gill 2020), despite being more statistically powered than that study. One possibility is that the original effect depended on the instantiation of the paradigm used, in which harming the passenger always involved taking a deliberate action (swerving into a barrier) whereas harming the pedestrian always involved inaction (staying on the existing path until hitting the pedestrian located in the vehicle's path). In contrast, the current studies presented a moral choice without specifying the type of action required. Harm from actions is typically viewed as more morally wrong than harm from inactions/omissions (Baron 1994; Folkes and Kamins 1999; Seggie, Griffith, and Jap 2013), because it is easier to coordinate punishment against clearly observable actions (De Freitas et al. 2019). Thus, participants in previous work may have exploited this fact in the AV (vs. HDV) condition, since they could both argue that they had no control of the AV and that the vehicle was headed for the pedestrian anyway. Deliberately swerving to harm the passenger may have seemed likely to attract blame; not to mention consumers were selfishly incentivized to avoid this option. Future work should investigate whether the original Gill (2020) findings hold when harming the *passenger* entails inaction, as when the AV is already headed for a barricade but can swerve to harm the pedestrian instead.

STUDY 5: UNSELFISH ALIBIS

Study 5 aims to arbitrate between whether consumers have consistently egalitarian preferences for AVs or only when their reputations are on the line. To this end, this study tests whether consumers act selfishly when provided with an unselfish alibi—a pretext that allows them to plausibly deny that they were acting selfishly. If they do not, then this provides strong evidence for egalitarianism, showing that consumers will reject selfish options even when the reputational repercussions are low. But if they do, then this suggests that the egalitarian choices found in the previous studies were largely driven by reputational concerns, rather than by an uncompromising concern for egalitarianism. Practically-speaking, it would suggest that unselfish alibis are an effective way to appeal to selfish preferences without inviting consumer backlash.

To arbitrate between these possibilities, participants were given a choice between an egalitarian option and a rule-based one (the unselfish alibi), which explained that in a dilemma the AV chooses whom to harm by driving into whatever is closest: if the barrier is closer than the pedestrian, then it drives into the barrier and kills the passenger; and vice versa if the pedestrian is closer than the barrier. Crucially, the study manipulated whether the AV happened to currently be facing a dilemma in which either the pedestrian was closer (creating a selfish temptation to choose the rule-based option) or the barrier was closer (creating no such temptation). I predicted that participants would be more likely to choose the rule-based option when tempted by the selfish outcome, because the unselfish alibi would allow them to plausibly defend their selfish choice (e.g., they could argue that the rule-based option seemed reasonable and fair in general, or that they did not realize the repercussions it would have in the current scenario).

Method

I aimed to collect 624 participants from Amazon Mechanical Turk and stopped collecting data when this request was filled by Amazon, resulting in 570 participants, who passed attention checks and completed the survey in exchange for $0.50. I excluded 252 participants who failed two challenging comprehension checks requiring close attention (described below), resulting in a final sample of 318 participants (M_{age} = 52, 39% female).

All participants were asked to imagine a scenario from the first-person perspective of ‘you’. They were assigned to one of two conditions, in which they chose between an egalitarian or rule-based algorithm. All participants read the following scenario and options:

> "Imagine that you are driving alone in an autonomous self-driving car that is confronted with the following dilemma: **Whether to drive into a barrier and kill the passenger in the car (you), or whether to drive into and kill a pedestrian.**
>
> The self-driving car can make its decision in one of two ways:
>
> - *Proximity rule:* It kills whoever is closest. If the barrier is closer than the pedestrian, it kills you. But if the pedestrian is closer than the barrier, it kills the pedestrian.
> - *No discrimination:* It randomly chooses whether to kill or save the passenger in the car or the pedestrian."

Then, depending on the condition, they were told one of the following:

> "In this particular case, the barrier is closer to the AV than the pedestrian."
> OR
> "In this particular case, the pedestrian is closer to the AV than the barrier."

Whereas choosing the proximity rule in the first case would kill the passenger (the prosocial outcome), choosing the same rule in the second case would kill the pedestrian instead (the selfish outcome). All participants chose how the AV should be programmed:

> **"The self-driving car should be programmed to…**
>
> - Use the proximity rule
> - Use no discrimination"

Next, they answered two comprehension checks about the proximity rule and details of the scenario that affected whether choosing the rule would result in a prosocial or selfish outcome:

> "According to the scenario, which of the following is true about the proximity rule?
>
> - The AV asks the driver to sit as close as possible to the steering wheel.
> - In a dilemma, the AV kills whoever is closest to it.
> - This is a trick question. Neither of the answers is correct.
>
> According to the scenario, which of the following was true?
>
> - The barrier was closer to the AV than the pedestrian.
> - The pedestrian was closer to the AV than the barrier.
> - This is a trick question. Neither of the answers is correct."

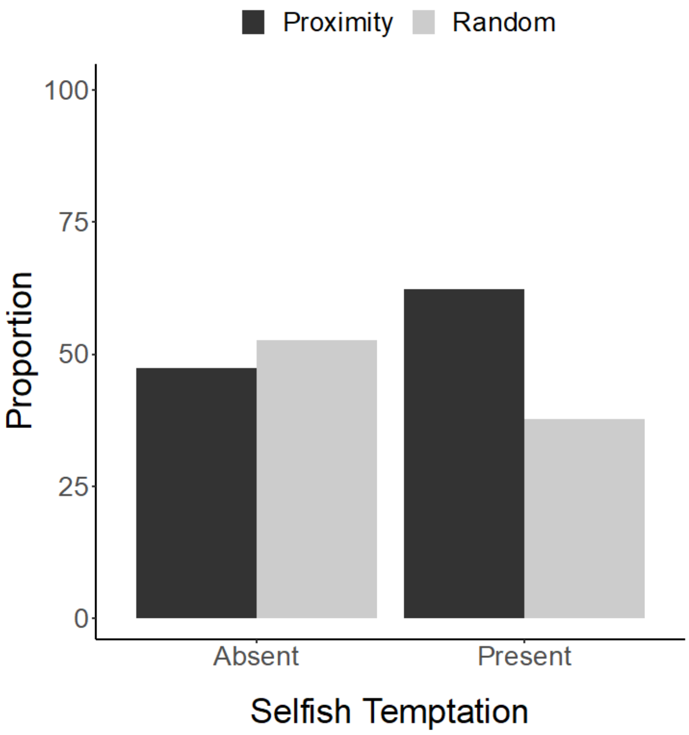
Since in this experiment it was crucial that participants understood both the proximity rule and the details of the scenario that the passenger was in, participants were excluded for answering either of the checks incorrectly. They then completed the demographics items.

Results

A logistic regression analysis found a main effect of whether consumers were tempted to choose selfishly (*b* = 0.60, *p* = .009; figure 5), with participants being more likely to choose the rule-based option when tempted by a selfish outcome than a prosocial outcome (Selfish: 62.24%, Prosocial: 47.43%)[2].

FIGURE 5

RESULTS IN STUDY 5



[2] Unsurprisingly, the results are not statistically significant if the analyses are repeated without exclusions (*b* = 0.26, *p* = 0.128), although they continue to be in the predicted direction, with more participants choosing the rule-based option when tempted by a selfish outcome than when not (Selfish: 55.16%, Prosocial: 48.79%).

Conclusion

Helping to arbitrate between whether consumers are truly egalitarian or selectively express egalitarian preferences when their reputations are at risk, study 5 found that consumers are more likely to choose selfishly when given a plausible pretext for doing so.

STUDY 6: UNSELFISH ALIBIS REPLICATION WITH ADVERTISEMENT

Study 6 tests whether the unselfish alibi effect generalizes to a more naturalistic advertisement setting. Participants are shown three advertisements with different slogans that are prosocial, overtly selfish, or ambiguously selfish (the unselfish alibi). I predict that they will be more likely to say they will ride with an AV brand that has an ambiguously selfish slogan than an overtly selfish one.

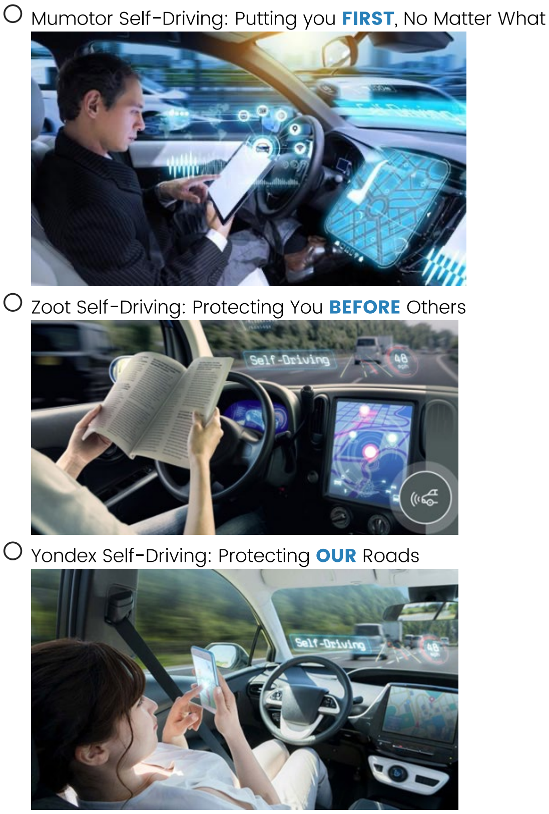
Method

I aimed to collect 312 participants from Amazon Mechanical Turk and stopped collecting data when this request was filled by Amazon, resulting in 308 participants, who passed attention checks and completed the survey in exchange for $0.50. I excluded 16 participants for failing two challenging comprehension checks requiring close attention (described below), resulting in a final sample of 292 participants.

Participants were asked to consider advertisements for three brands, which had slogans that were prosocial (“Protecting OUR Roads”), overtly selfish (“Protecting You BEFORE Others”), or ambiguously selfish (“Putting You First, No Matter What”) (see figure 6). The third statement is ambiguously selfish because “no matter what” is a vaguer, generic statement that is open to several interpretations (e.g., putting the customer before the company’s other interests, or transporting the customer even in adverse weather and traffic conditions). Participants were then asked to choose one vehicle they would ride in: *Based on these ads, please choose which vehicle you would be most willing to ride in?* The pairing of advertising slogans with images was counterbalanced between-subjects, and the order of the three ads on the page was randomized within-subjects.

FIGURE 6

STIMULI IN STUDY 6



○ Mumotor Self-Driving: Putting you **FIRST**, No Matter What

○ Zoot Self-Driving: Protecting You **BEFORE** Others

○ Yondex Self-Driving: Protecting **OUR** Roads

NOTE.— Sources from top to bottom:

https://www.cprime.com/wp-content/uploads/2022/07/archer_blog_security_testing_for_self-driving_cars__5_best_practices_cprimestudios_image.jpeg https://piernext.portdebarcelona.cat/wp-content/uploads/2021/05/f3-1000x500.jpg

https://images.newscientist.com/wp-content/uploads/2017/10/13174202/gettyimages-829192848.jpg?crop=16:9,smart&width=1200&height=675&upscale=true

Next, as a manipulation check of whether the items truly differed in perceived selfishness, participants answered several questions about each ad, on separate pages for each ad (with page order counterbalanced between-subjects). For each ad, they were shown the slogan again and asked questions about whether others would think the participant was selfish for choosing that option. In order to get at the extent to which the choice would be considered a conspicuous violation of egalitarianism, participants were asked about not just whether others would privately think that the participant was selfish (private knowledge), but also whether they would think the participant knew they thought this (shared knowledge) and whether they thought it would be commonly known by both of them (common knowledge; De Freitas et al. 2019) :

> "Imagine that you choose to ride in this vehicle on a regular basis, and a friend of yours discovers this. Please rate the extent to which you agree with the following statements (1=completely disagree, 100=completely agree):
>
> Your friend would think that the real reason you chose the vehicle is to selfishly protect yourself.
>
> Your friend would believe that you know that they think the real reason you picked the vehicle is to selfishly protect yourself.
>
> You and your friend would both think that you both think that the real reason you picked the vehicle is to selfishly protect yourself."

Participants then answered two comprehension checks about what they saw and what the questions were about:

> "According to the scenario, which of the following is true about the images you saw?
> - They were advertisements for different self-driving cars.
> - They were advertisements for different regular (non self-driving) cars.
> - This is a trick question. Neither of the above options is correct.

What were you asked to imagine in the questions?

- That you were selected as a finalist for a competition.
- That your friend discovered what option you chose.
- This is a trick question. Neither of the above options is correct."

Participants were excluded for failing one or more of these checks. I accidentally neglected to collect demographic information.

Results

Participant's choices were significantly affected by the advertisement slogans X^2 (1, N = 292) = 37.86, $p < .001$. The most chosen option was prosocial (44.9%), followed by ambiguously selfish (38.4%), then overtly selfish (16.8%). Importantly, consumers were more likely to choose the ambiguously selfish option than the overtly selfish one, consistent with the unselfish alibi hypothesis, X^2 (1, N = 292) = 24.65, $p < .001$.

Consistent with the idea that choosing the overtly selfish option constituted a conspicuous violation of the egalitarian norm, participants were more likely to say that others would view them as selfish for choosing the overtly selfish option—whether this was primary, shared, or common knowledge; although the effect for common knowledge was marginal (see Table 1).

TABLE 1

AGREEMENT THAT CHOOSING ADVERTISEMENT IS VIEWED AS SELFISH

	Prosocial Ad	Unselfish Alibi Ad	Selfish Ad	Selfish v Alibi
Primary Knowledge	23.1 (22.0)	52.6 (31.0)	56.9 (30.5)	$t(291)=3.16^{**}$
Shared Knowledge	21.2 (21.0)	51.7 (30.9)	54.8 (30.7)	$t(291)=2.24^{*}$
Common Knowledge	21.6 (22.1)	51.6 (31.3)	54.2 (30.8)	$t(291)=1.91^{+}$

NOTE.—$^{+}$ $p < .10$, * $p < .05$, ** $p < .01$

Discussion

Study 6 found further evidence for the unselfish alibi effect in a more naturalistic advertisement context. The egalitarian option was also very popular, consistent with the previous findings that consumers choose this option when it is explicitly proffered.

STUDY 7: NUDGING CONSUMERS TOWARD PROSOCIALITY

While appealing to selfish instincts may benefit individual consumers and AV brands, it is not necessarily conducive to social coordination on roads, which relies on AVs following similar social norms rather than acting selfishly (Fehr and Schurtenberger 2018). Further, implying to consumers that AVs preferentially 'choose' passengers over pedestrians is a misrepresentation of how they are programmed, which is to avoid hitting anyone regardless of their social characteristics (Censi et al. 2019; De Freitas et al. 2021).

More broadly, in some cases policymakers and competitors may want to discourage selfish choices. Thus, study 7 investigates whether it is possible for an unselfish choice to win out against unselfish alibis. I reasoned that consumers would rather not make the taxing tradeoff between selfishness and prosociality unless deemed necessary (Leonhardt, Keller, and Pechmann 2011; Lin and Reich 2018). Thus, they should be more likely to make unselfish choices when given choice sets that implicitly undermine the need to make a tradeoff in the first place. Here, I pit unselfish alibis against the option of 'minimizing harm', where instead of the AV choosing whom to harm in a dilemma it does its best to minimize harm to all parties.

Method

I aimed to collect 624 participants from Prolific and stopped collecting data when this request was filled by Amazon, resulting in 597 participants, who passed attention checks and completed the survey in exchange for $0.67. I excluded 299 participants who failed two challenging comprehension checks requiring careful attention (described below), resulting in a final sample of 298 participants (M_{age} = 36, 51% female).

The design was like study 5, except this time the scenario was set up so that the participant was always selfishly tempted to pick the unselfish alibi (the 'proximity rule' would favor saving the passenger). I manipulated whether participants were given a dichotomous forced choice between the proximity rule and no discrimination (as in study 5), or an additional third option involving harm minimization. For instance, after reading about the dilemma, participants in the Three Choice condition chose from the following options:

> "The self-driving car can make its decision in one of three ways:
>
> *1. Proximity rule*: It kills whoever is closest to the AV. If the barrier is closer than the pedestrian, it kills you. But if the pedestrian is closer than the barrier, it kills the pedestrian.
> *2. No discrimination:* It randomly chooses whether to kill or save the passenger in the car or the pedestrian.
> *3. Minimize death*: Instead of choosing whom to kill or save, it minimizes any death or risk of death.
>
> In this particular case, the pedestrian is closer to the AV than the barrier.
>
> **The self-driving car should be programmed to...**"

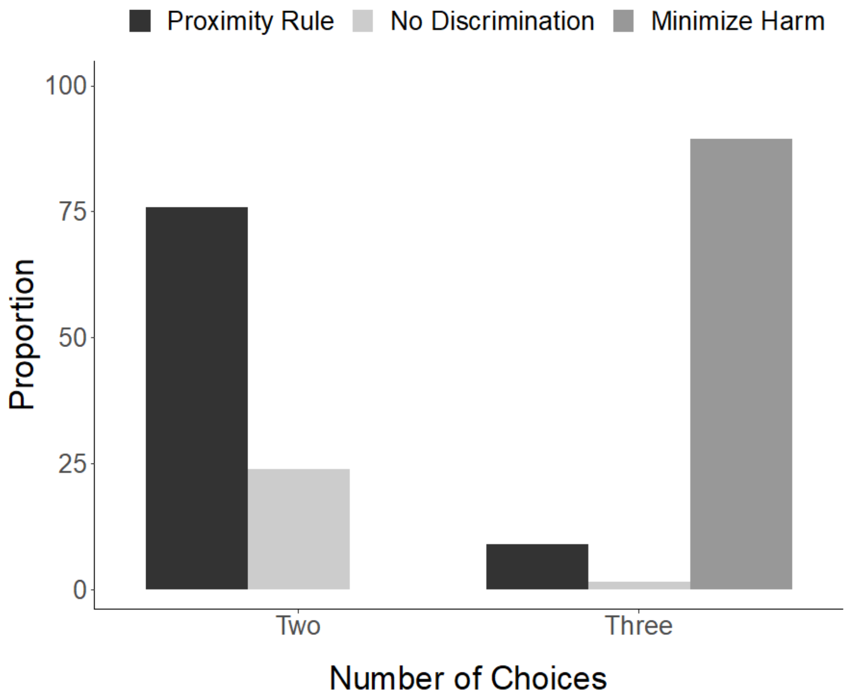
Participants then answered the same comprehension checks as in study 5, followed by demographics items. As in study 5, it was important that participants understood the proximity rule and the temptation to choose it, so I excluded participants for failing any of the checks.

Results

To test the effect of Choice Type on whether consumers selfishly choose the unselfish alibi (the proximity rule), I created a dichotomized choice variable coded as ‘1’ if consumers selfishly chose the proximity rule and ‘0’ otherwise. A logistic regression analysis found that these dichotomized choices were predicted by Choice Type ($b = 3.47$, $p < .001$). As predicted, most participants (76%) chose the proximity rule when given two options, whereas only 9% chose the proximity rule when additionally offered the harm minimization option. Instead, most participants (89%) chose the harm minimization option when offered it (figure 7).

FIGURE 7

RESULTS IN STUDY 7



Discussion

Participants tended to choose selfishly when given a dichotomous choice set between an unselfish alibi and no discrimination, but not when additionally offered a third option to minimize harm—perhaps because this option implicitly undermined the need for a dichotomy in the first place. This result suggests that consumers are not irredeemably selfish but can be encouraged via choice set framing to make prosocial choices. Practically, policymakers and competitors can use such interventions to encourage unselfish choices that are more aligned with how the technology works and better for society.

GENERAL DISCUSSION

Across seven studies, I find that consumer preferences for AVs are best described as selfish. At the same time, consumers do not want to appear as though they are willing to violate the egalitarian norm. Studies 1-2 find that they express outrage when the norm is deliberately (but not indiscriminately) violated, and studies 3-4 find that they will pick egalitarian settings when explicitly proffered. Tellingly, however, selfish options disguised as unselfish are chosen more frequently than similar options without a selfish temptation (studies 5-6), showing that consumers are strategic in when they choose selfishly. Providing some optimism for regulators and competitors wishing to counteract such effects, consumers can be encouraged to choose unselfishly when offered choice sets that undermine the need to make a tradeoff (study 7).

These findings add nuance to previous work on consumer moral preferences for AVs. While the findings cohere with previous findings of selfish preferences (Bonnefon et al. 2016; Frank et al. 2019; Gill 2020), they show that consumers are highly strategic about when they act on these selfish instincts. Indeed, in cases where their reputations are at stake, consumers make choices that appear highly egalitarian. Second, this work contributes to literature on choice sets, most of which has studied how choice sets affect what kinds of information consumers do or do not notice (Bazerman et al. 1992; De Freitas and Johnson 2018; Hsee et al. 1999). The current findings suggest that choice sets also send differential social signals about a person's character if certain options made available in the set are rejected, damaging reputations. Third, this work extends previous work on plausible deniability in consumption, which has focused on how consumers justify buying attractive items without feeling guilty, as in 'functional alibis' (Keinan et al. 2016) and 'accidental product breakages' (Bellezza et al. 2017; Shani et al. 2020). The present studies add the notion of unselfish alibis, where the alibi is used to justify resolving a selfish-prosocial tradeoff in favor of selfishness. What all phenomena have in common is that the consumer is trying to maintain a desirable reputation, yet the exact tradeoff differs across domains and is arguably more societally consequential in the current context.

Limitations and Future Directions

Most of the scenarios in these studies involved selfish-prosocial tradeoffs with extreme consequences. Partially, this was because consumers are concerned about such cases in the domain of AVs, and because this work revisited previous findings employing such dilemmas. Yet future work should test whether consumers show similar response patterns for mundane tradeoffs, such as performing a rolling stop in order to reach a destination earlier. As another relevant example, in January of 2022 the automotive firm, Tesla, started offering consumers the ability to choose whether its semi-autonomous software drove itself in "chill", "average" or "assertive" mode (Roth 2022). It seems likely that the indirect language of "assertive" created an unselfish alibi, allowing consumers to plausibly justify choosing this option (e.g., "I like a confident driver", or "driving assertively is actually safer"). Future work should explore consumer reactions to such settings, both in the field and the lab. Given the social pressure to appear egalitarian, such work should also aim to measure consumer behavior directly, such as by measuring when and why Tesla customers use assertive mode. A final question is whether consumers are aware that they resolve tradeoffs in a selfish direction (Nisbett and Wilson 1977) and are susceptible to unselfish alibis.

Practical Implications

The findings provide a warning and recommendation for firms wishing to appeal to consumer's selfish instincts. First, firms should not do this overtly, even when targeting their own customers and prospects. Second, if they provide admissions of selfish outcomes, they are better off making these seem unintentional rather than deliberate. Third, they can successfully appeal to selfish preferences while avoiding consumer backlash by employing unselfish alibis.

As for regulators and competitors wishing to discourage selfish choices, they can communicate or implicitly suggest that making selfish-prosocial tradeoffs is unnecessary, given that consumers would rather avoid making these challenging tradeoffs if deemed unnecessary.

More broadly, our findings suggest a fundamental tension between the vision of safe AVs on the one hand and consumer's selfish preferences on the other. The question is to what extent these selfish preferences will be allowed to jeopardize the safety promises of this technology. For instance, on February of 2022, a month after Tesla's assertive mode was deployed, it was forced to recall 53,8222 of its vehicles because the rolling stop behavior was deemed dangerous (McFarland 2022). Yet, a year later, the driving modes are available again in some form. As more AVs are increasingly deployed on our roads, how will the tug of war between selfish settings and safety play out?

REFERENCES

Allison, Scott T and David M Messick (1990), "Social Decision Heuristics in the Use of Shared Resources," *Journal of Behavioral Decision Making*, 3 (3), 195-204.

Andreoni, James and Ragan Petrie (2004), "Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising," *Journal of public Economics*, 88 (7-8), 1605-23.

Baron, Jonathan (1994), "Nonconsequentialist Decisions," *Behavioral and Brain Sciences*, 17 (1), 1-10.

Bazerman, Max H, George F Loewenstein, and Sally Blount White (1992), "Reversals of Preference in Allocation Decisions: Judging an Alternative Versus Choosing among Alternatives," *Administrative Science Quarterly*, 220-40.

Bazerman, Max H, Sally Blount White, and George F Loewenstein (1995), "Perceptions of Fairness in Interpersonal and Individual Choice Situations," *Current Directions in Psychological Science*, 4 (2), 39-43.

Bechwati, Nada Nasr and Maureen Morrin (2003), "Outraged Consumers: Getting Even at the Expense of Getting a Good Deal," *Journal of Consumer Psychology*, 13 (4), 440-53.

Bellezza, Silvia, Joshua M Ackerman, and Francesca Gino (2017), "“Be Careless with That!” Availability of Product Upgrades Increases Cavalier Behavior toward Possessions," *Journal of Marketing Research*, 54 (5), 768-84.

Bereczkei, Tamas, Bela Birkas, and Zsuzsanna Kerekes (2007), "Public Charity Offer as a Proximate Factor of Evolved Reputation-Building Strategy: An Experimental Analysis of a Real-Life Situation," *Evolution and Human behavior*, 28 (4), 277-84.

Bigman, Yochanan E and Kurt Gray (2020), "Life and Death Decisions of Autonomous Vehicles," *Nature*, 579 (7797), E1-E2.

Bonnefon, Jean-François, Azim Shariff, and Iyad Rahwan (2016), "The Social Dilemma of Autonomous Vehicles," *Science*, 352 (6293), 1573-76.

Censi, Andrea, Konstantin Slutsky, Tichakorn Wongpiromsarn, Dmitry Yershov, Scott Pendleton, James Fu, and Emilio Frazzoli (2019), "Liability, Ethics, and Culture-Aware Behavior Specification Using Rulebooks," *2019 International Conference on Robotics and Automation (ICRA)*, 8536-42.

Chance, Zoë and Michael I Norton (2015), "The What and Why of Self-Deception," *Current Opinion in Psychology*, 6, 104-07.

Crockett, Molly J (2017), "Moral Outrage in the Digital Age," *Nature Human Behaviour*, 1 (11), 769–71.

Dana, Jason, Daylian M Cain, and Robyn M Dawes (2006), "What You Don’t Know Won’t Hurt Me: Costly (but Quiet) Exit in Dictator Games," *Organizational Behavior and Human Decision Processes*, 100 (2), 193-201.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang (2007), "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 33 (1), 67-80.

De Freitas, Julian, Sam E Anthony, Andrea Censi, and George Alvarez (2020), "Doubting Driverless Dilemmas," *Perspectives on Psychological Science*, 15, 1284-88.

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.

De Freitas, Julian and Samuel GB Johnson (2018), "Optimality Bias in Moral Judgment," *Journal of Experimental Social Psychology*, 79, 149-63.

De Freitas, Julian, Kyle Thomas, Peter DeScioli, and Steven Pinker (2019), "Common Knowledge, Coordination, and Strategic Mentalizing in Human Social Life," *Proceedings of the National Academy of Sciences*, 116 (28), 13751-58.

Delton, Andrew W, Max M Krasnow, Leda Cosmides, and John Tooby (2011), "Evolution of Direct Reciprocity under Uncertainty Can Explain Human Generosity in One-Shot Encounters," *Proceedings of the National Academy of Sciences*, 108 (32), 13335-40.

Deutsch, Morton (1975), "Equity, Equality, and Need: What Determines Which Value Will Be Used as the Basis of Distributive Justice?," *Journal of Social Issues*, 31 (3), 137-49.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey (2015), "Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err," *Journal of Experimental Psychology: General*, 144 (1), 114-26.

Fehr, Ernst and Ivo Schurtenberger (2018), "Normative Foundations of Human Cooperation," *Nature Human Behaviour*, 2 (7), 458-68.

Fisher, Robert J and David Ackerman (1998), "The Effects of Recognition and Group Need on Volunteerism: A Social Norm Perspective," *Journal of Consumer Research*, 25 (3), 262-75.

Folkes, Valerie S and Michael A Kamins (1999), "Effects of Information About Firms' Ethical and Unethical Actions on Consumers' Attitudes," *Journal of Consumer Psychology*, 8 (3), 243-59.

Ford, Brett Q, Matthew Feinberg, Phoebe Lam, Iris B Mauss, and Oliver P John (2018), "Using Reappraisal to Regulate Negative Emotion after the 2016 Us Presidential Election: Does Emotion Regulation Trump Political Action?," *Journal of Personality and Social Psychology*, 117 (5), 998-1015.

Frank, Darius-Aurel, Polymeros Chrysochou, Panagiotis Mitkidis, and Dan Ariely (2019), "Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles," *Scientific Reports*, 9 (1), 1-19.

Gershon, Rachel, Cynthia Cryder, and Leslie K John (2020), "Why Prosocial Referral Incentives Work: The Interplay of Reputational Benefits and Action Costs," *Journal of Marketing Research*, 57 (1), 156-72.

Gill, Tripat (2020), "Blame It on the Self-Driving Car: How Autonomous Vehicles Can Alter Consumer Morality," *Journal of Consumer Research*, 47 (2), 272-91.

Gino, Francesca, Michael I Norton, and Roberto A Weber (2016), "Motivated Bayesians: Feeling Moral While Acting Egoistically," *Journal of Economic Perspectives*, 30 (3), 189-212.

Griskevicius, Vladas and Douglas T Kenrick (2013), "Fundamental Motives: How Evolutionary Needs Influence Consumer Behavior," *Journal of Consumer Psychology*, 23 (3), 372-86.

Hardy, Charlie L and Mark Van Vugt (2006), "Nice Guys Finish First: The Competitive Altruism Hypothesis," *Personality and Social Psychology Bulletin*, 32 (10), 1402-13.

Hayes, Andrew F. (2012), "Process: A Versatile Computational Tool for Observed Variable Mediation, Moderation, and Conditional Process Modeling [White Paper]," Retrieved from http://www.afhayes.com/public/process2012.pdf.

Hsee, Chrisopher K, George F Loewenstein, Sally Blount, and Max H Bazerman (1999), "Preference Reversals between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis," *Psychological Bulletin*, 125 (5), 576.

Jordan, Jillian J and David G Rand (2020), "Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment in One-Shot Anonymous Interactions.," *Journal of Personality and Social Psychology*, 118 (1), 57-88.

Keinan, Anat, Ran Kivetz, and Oded Netzer (2016), "The Functional Alibi," *Journal of the Association for Consumer Research*, 1 (4), 479-96.

Khatoon, Sajira and Varisha Rehman (2021), "Negative Emotions in Consumer Brand Relationship: A Review and Future Research Agenda," *International Journal of Consumer Studies*, 45 (4), 719-49.

Leonhardt, James M, L Robin Keller, and Cornelia Pechmann (2011), "Avoiding the Risk of Responsibility by Seeking Uncertainty: Responsibility Aversion and Preference for Indirect Agency When Choosing for Others," *Journal of Consumer Psychology*, 21 (4), 405-13.

Li, T and L Cheer (2016), "Mercedes-Benz Admits Automated Driverless Cars Would Run over a Child Rather Than Swerve and Risk Injuring the Passengers Inside," in *Daily Mail*.

Lin, Stephanie C and Taly Reich (2018), "To Give or Not to Give? Choosing Chance under Moral Conflict," *Journal of Consumer Psychology*, 28 (2), 211-33.

Lindenmeier, Jörg, Christoph Schleer, and Denise Pricl (2012), "Consumer Outrage: Emotional Reactions to Unethical Corporate Behavior," *Journal of Business Research*, 65 (9), 1364-73.

Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge (2019), "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research*, 46 (4), 629-50.

McFarland, Matt (2016), "Tesla's Autopilot Probed by Government after Crash Kills Driver," *URL http://money. cnn. com/2016/06/30/technology/tesla-autopilotdeath.*

——— (2022), "Tesla Recalls All 53,822 Vehicles with Its 'Full Self-Driving' Fetaure," https://www.cnn.com/2022/02/01/cars/tesla-fsd-stop-sign/index.html.

Messick, David M and Terry Schell (1992), "Evidence for an Equality Heuristic in Social Decision Making," *Acta Psychologica*, 80 (1-3), 311-23.

Morris, DZ (2016), "Mercedes-Benz's Self-Driving Cars Would Choose Passenger Lives over Bystanders," in *Fortune*.

Nader, Ralph (1965), *Unsafe at Any Speed: The Designed-in Dangers of the American Automobile*: New York: Grossman.

Neuberg, Steven L, Douglas T Kenrick, and Mark Schaller (2011), "Human Threat Management Systems: Self-Protection and Disease Avoidance," *Neuroscience & Biobehavioral Reviews*, 35 (4), 1042-51.

Nisbett, Richard E and Timothy D Wilson (1977), "Telling More Than We Can Know: Verbal Reports on Mental Processes," *Psychological Review*, 84 (3), 231.

Novak, Thomas P (2020), "A Generalized Framework for Moral Dilemmas Involving Autonomous Vehicles: A Commentary on Gill," *Journal of Consumer Research*, 47 (2), 292-300.

Nowak, Martin A and Karl Sigmund (2005), "Evolution of Indirect Reciprocity," *Nature*, 437 (7063), 1291.

Roth, Emma (2022), "Tesla's 'Full Self-Driving' Beta Has an 'Assertive' Driving Mode That 'May Perform Rolling Stops'."

Salerno, Jessica M and Liana C Peter-Hagene (2013), "The Interactive Effect of Anger and Disgust on Moral Outrage and Judgments," *Psychological Science*, 24 (10), 2069-78.

Sampson, Edward E (1975), "On Justice as Equality," *Journal of Social Issues*, 31 (3), 45-64.

Seggie, Steven H, David A Griffith, and Sandy D Jap (2013), "Passive and Active Opportunism in Interorganizational Exchange," *Journal of Marketing*, 77 (6), 73-90.

Shani, Yaniv, Gil Appel, Shai Danziger, and Ron Shachar (2020), "When and Why Consumers "Accidentally" Endanger Their Products," *Management Science*, 66 (12), 5757-82.

Singh, Santokh (2015), "Critical Reasons for Crashes Investigated in the National Motor Vehicle Crash Causation Survey," https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115.

Soetevent, Adriaan R (2005), "Anonymity in Giving in a Natural Context—a Field Experiment in 30 Churches," *Journal of public Economics*, 89 (11-12), 2301-23.

Taylor, M (2016), "Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians," in *Car and Driver*.

Tetlock, Philip E (2003), "Thinking the Unthinkable: Sacred Values and Taboo Cognitions," *Trends in Cognitive Sciences*, 7 (7), 320–24.

Vijayenthiran, V (2016), "Mercedes Is Backtracking on Claims Its Self-Driving Cars Will Kill Pedestrians over Passengers in Close Calls," in *Businessinsider*.

Wakabayashi, Daisuke (2018), "Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam," *The New York Times*, 19 (03).

Welle, Ben, Anna Bray Sharpin, Claudia Adriazola-Steil, Soames Job, Marc Shotten, Dipan Bose, Amit Bhatt, Saul Alveano, Marta Obelheiro, and Tolga Imamoglu (2018), "Sustainable & Safe: A Vision and Guidance for Zero Road Deaths," https://www.wri.org/publication/sustainable-and-safe-vision-and-guidance-zero-road-deaths.

Wilmot, Stephen (2023), "You May Be Able to Buy a Self-Driving Car after All," https://www.wsj.com/articles/you-may-be-able-to-buy-a-self-driving-car-after-all-11673018475.

Wu, Junhui, Daniel Balliet, and Paul AM Van Lange (2016), "Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation," *Scientific Reports*, 6 (1), 1-8.