Article

Social Psychological and
Personality Science
1-8
© The Author(s) 2018
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1948550618755875
journals.sagepub.com/home/spp
SAGE

Autonomous Vehicles and the Attribution of Moral Responsibility

Ryan M. McManus[1] and Abraham M. Rutchick[1]

Abstract

With the imminent advent of autonomous vehicles (AVs) comes a moral dilemma: How do people assign responsibility in the event of a fatal accident? AVs necessarily create conditions in which "drivers" yield agency to a machine. The current study examines how people make attributions of blame and praise in this context. Varying the features of AV technology affected how responsible a "driver" (who purchased the vehicle) is perceived to be following a deadly crash. The findings provide support for agency and commission as crucial bases of moral judgment. They also raise questions about how morally contradictory actions are perceived and underscore the need for research examining how moral responsibility is distributed among multiple potentially culpable agents. Pragmatically, these findings suggest that regulating (or declining to regulate) how AVs are programmed may strongly influence perceptions of moral and legal culpability.

Keywords

autonomous vehicle, morality, agency, blame, praise

Autonomous (i.e., self-driving) vehicles (AVs) have the potential to reduce traffic accidents by up to 90% (Gao, Hensley, & Zielke, 2014) and will likely be commercially available by 2021 (Boudette & Markoff, 2016). One challenge in developing AVs is determining their moral programming (Matyszczyk, 2016). If the vehicle approaches a potentially lethal situation, should it be programmed to maximize the possible number of lives saved or should it maneuver to save its occupant(s) at all costs? If no life is more valuable than another (Bentham, 1879/1983; Mill, 1863; Singer, 1972), maximizing the good means maximizing the number of lives saved, even if the vehicle's occupant(s) die. This utilitarian programming could, however, conflict with deontic principles (i.e., individual rights and autonomy; Kant, 1797/2002; Scanlon, 1998).

While the programming problem highlights the need for practical answers to age-old problems in moral philosophy, this scenario also raises an additional concern regarding moral and legal culpability: How much fault lies with the owner of the vehicle? For example, if an owner had a hand in the vehicle's programming, should they be held *more* responsible than if the programming had been independently determined by the manufacturer or a third-party programmer? Furthermore, this technology will afford people an opportunity to outsource their future moral decision-making abilities. How will one's decision to outsource their moral agency affect judgments of responsibility? The current article investigates attributions of blame and praise in scenarios in which fatal accidents involve AVs.

Consider the following scenario, the "tunnel problem" (Millar, 2014):

> You are travelling along a single lane mountain road in an autonomous car that is fast approaching a narrow tunnel. Just before entering the tunnel, a child attempts to run across the road but trips in the center of the lane, effectively blocking the entrance to the tunnel. The car has but two options: hit and kill the child, or swerve into the wall on either side of the tunnel, thus killing you. How should the car react?

An opinion poll was conducted by the Open Roboethics Initiative (2014) in which this scenario was presented to 113 people; 64% believed that the vehicle should continue straight, killing the child. These results mirror recent research demonstrating that people generally do not believe AVs should sacrifice their occupant if only one life could be saved by doing so (Bonnefon, Shariff, & Rahwan, 2016). When asked about who should make that decision, 44% believed that the choice should lie with the passenger/consumer, while 33% thought it should be left to lawmakers, and 12% reported it should be up to the manufacturer/designer (Open Roboethics Initiative, 2014). Bonnefon, Shariff, and Rahwan (2016) also showed that people approve of utilitarian programming for others' AVs but would themselves prefer vehicles that protect them at all costs.

[1] Department of Psychology, California State University, Northridge, CA, USA

Corresponding Author:
Ryan M. McManus, Department of Psychology, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330, USA.
Email: ryanmc1289@gmail.com

Given the extensive literature on self-serving biases (e.g., Campbell & Sedikides, 1999; Sedikides & Alicke, 2012), it is not surprising that people value their own lives over others'. Although such accidents will presumably be rare, they are likely to garner broad public coverage due to their emotional impact (Bonnefon et al., 2016). Therefore, the study of outsourced moral agency in the context of AVs has both theoretical import and practical relevance. To understand, though, how moral culpability may change depending on the regulation of programming or the amount of control the owner has when programming is unregulated, it is important to consider research relevant to agency and intentionality.

It has been demonstrated that people have a cognitive template—a moral dyad—that automatically categorizes people into agents (who act with intent) and patients (who experience suffering) when harm is perceived (Gray, Schein, & Ward, 2014; Gray, Waytz, & Young, 2012). The more "dyadic" a transgression—the clearer it is that there is an intentional agent, vulnerable patient, and causation of damage—the more it is perceived as harmful. The more harmful a transgression, the more it is perceived as immoral (Gray & Schein, 2017). Although this model explains many findings in the moral judgment literature, it is unclear how such judgments will unfold in the context of AVs; by definition, AVs create conditions in which intentional agents are either absent or a significant period of time has passed since the crucial decision (i.e., the purchase or programming of the vehicle).

Relatedly, Pizarro, Uhlmann, and Bloom (2003) demonstrated that judgments of responsibility are attenuated when the manner in which an event occurred deviates from the expected causal chain. For example, if a transgressor intends to stab someone in the street but is knocked into their target by a passing jogger (causing them to stab their target anyway), blame is attenuated when compared to the case in which the transgressor carries out their stabbing in a causally normal way. Pizarro, Uhlmann, and Bloom (2003) argued that this attenuation is driven by lack of "intention-in-action." Searle (1983) stated that an intentional action consists of two components: a prior intention (the deliberate planning) and an intention-in-action (the direct mental cause of an action). Findings from Pizarro, Uhlmann, and Bloom (2003) support the application of this concept to judgments of responsibility. When an agent has a prior intention and an intention-in-action (i.e., in causally normal cases), more responsibility is assigned than if the latter intention is absent at the critical moment (i.e., in causally deviant cases). This is particularly relevant to accidents involving programmable AVs; even if consumers purchase vehicles with the intent to save (or sacrifice) themselves in crash scenarios, it is unclear if such intentions would be perceived as being present at the time of the accident.

Applying this evidence to the fatal crash scenarios considered in the current study, we anticipated that blame and praise would vary according to the agency of the drivers. As drivers who manually control their vehicles should be perceived as more agentic than "drivers" of AVs that are preprogrammed by the manufacturer, following an accident in which the driver is saved but others die (a nonutilitarian/selfish outcome), we predicted that drivers of manual vehicles would be blamed more than drivers of manufacturer-programmed AVs. Similarly, drivers who are sacrificed to save several lives (a utilitarian/selfless outcome) should be praised more in manual vehicles than in manufacturer-programmed AVs.

Moral judgments also depend on the distinction between commission and omission (Baron & Ritov, 1994; Greene et al., 2009; Kordes-de Vaal, 1996). Spranca, Minsk, and Baron (1991) demonstrated that harmful commissions (e.g., lying to the police to protect a friend) are judged as more blameworthy than harmful omissions (e.g., saying nothing when a guilty friend is mistakenly cleared). Therefore, we hypothesized that, after selfish outcomes, drivers who actively choose a selfish algorithm (active/commissive harm) would be blamed more than those for whom the algorithm was programmed by the manufacturer (passive/omissive harm). Similarly, for selfless outcomes, drivers who choose the selfless algorithm should be praised more than those whose vehicles were programmed by the manufacturer. Although asymmetries exist in the way people attribute moral blame and praise when manipulating the impulsivity of an action (Pizarro, Uhlmann, & Salovey, 2003), it is unclear whether they differ along other important dimensions (Bartels, Bauman, Cushman, Pizarro, & McGraw, 2015). Thus, we hypothesize that manipulating agency will induce similar changes in blame and praise.

People are generally motivated to be consistent in their judgments and behavior (Festinger, 1957; Friedrich, 2002; Swann, Griffin, & Predmore, 1987). Given this, we predict that blame and praise will increase when drivers follow through on initial commitments (by declining to pull an "override switch" that allows the driver to take control). This amplification could also result from the perception that drivers have the ability to override their initial intentions but decline to do so (Buckwalter & Turri, 2015).

When a driver *does* use an override switch, it is unclear how they will be judged. Perhaps overriding an initial programming choice at the critical moment will be perceived as an impulsive, uncontrollable reaction leading to attenuated judgments for blame but not praise (Pizarro, Uhlmann, & Salovey, 2003). When a driver uses an override switch, they will have engaged in two distinct and morally contradictory actions (actively choosing a selfish or selfless program and then manually overriding it). It is possible that perceptions of the two actions cancel one another out, thus attenuating judgments of both blame and praise. Additionally, this scenario allows drivers to recapture their previously outsourced moral agency. Current theories of responsibility cannot easily explain this situation, as most research focuses on how responsibility is assessed following isolated incidents.

However, research in the negotiation literature suggests a counterintuitive possibility. Harford and Solomon (1967) demonstrated, using a Prisoner's Dilemma paradigm, that people are more trusting of a counterpart who initially acts noncooperatively and then becomes cooperative (a "reformed sinner") than a counterpart that is cooperative from the outset.

Conversely, participants were less trusting of a counterpart who initially cooperates and subsequently becomes noncooperative (a "lapsed saint") than one who is noncooperative from the outset. It is possible that this effect is driven by having more data points to make judgments about the counterpart's moral character. Some researchers argue that judgments of character carry important social information that judgments of actions do not (Pizarro & Tannenbaum, 2011; Uhlmann, Pizarro, & Diermeier, 2015), and in particular, that a person's moral character is a signal as to whether they should be trusted or avoided (Bigman & Tamir, 2016). When observers make judgments of a *person* who has performed an act, it has been demonstrated that they are using that information as a proxy for the moral integrity of the actor (Tannenbaum, Uhlmann, & Diermeier, 2011). Additionally, others' moral judgments and actions are perceived to reflect their trustworthiness (Everett, Crockett, & Pizarro, 2016). Perhaps, when judging someone who first makes a selfless moral commitment but at the critical moment selfishly overrides it, blame is amplified via increased distrust of the actor (i.e., he or she is someone who, if given the opportunity, will violate the norm of reciprocal altruism). Conversely, when judging someone who first makes a selfish commitment but selflessly overrides it, attributions of praise may be amplified because the actor has shown that they can overcome their baser impulses and sacrificed for the greater good. Accordingly, we predict that drivers who make a selfless choice but then use an override switch (allowing them to manually control the vehicle) to reverse that choice at the critical moment should be blamed more than drivers who make a selfish choice from the outset. Similarly, drivers who make, then override, a selfish choice should be praised more than drivers who make the selfless choice from the outset.

Method

Participants were recruited and compensated US$0.50 via Amazon's Mechanical Turk. An initial sample size of 220 was chosen to yield a minimum of 200 participants. This sample size affords a power of .80 to detect matched-pair comparison effects of $d = 0.20$, per G*Power 3.1(Faul, Erdfelder, Lang, & Buchner, 2007); these simple effects tests were the preregistered analyses that examined our specific hypotheses (i.e., comparing two vignettes to one another). After collecting this initial sample, the preregistered exclusion criterion was applied (providing a specific response that was contained in a lengthy vignette; there were two such attention checks), yielding only 153 participants who passed at least one check and 97 participants who passed both checks. Another sampling wave, consisting of an additional 140 participants, was conducted to ensure the desired sample size. Thus, a total of 360 people were surveyed (yielding, ultimately, 246 participants who passed at least one check and 158 who passed both checks).

Participants were instructed to imagine a time in the near future when roads would be shared by both autonomous and non-AVs. They then evaluated 10 hypothetical driving scenarios in randomized order. In each vignette, a consumer purchased a vehicle (the features of which varied across conditions) and months later faced a trolley-like dilemma while driving. He then either swerved, crashing into a cement truck that blocked an exit ramp (the selfless outcome), or did not swerve, striking two construction workers (the selfish outcome). As Bonnefon and colleagues (2016) demonstrated that the majority of people did not think AVs should sacrifice their occupants when only one life could be saved by doing so, it was decided to have more than one potential victim.

As noted, the nature of the vehicle's control system was manipulated. In one condition, the vehicle was controlled like a traditional car (the Manual Condition). In the other conditions, the vehicle was autonomous. The vehicle's programming was either chosen by the consumer when he bought it (the Driver-Programmed condition) or programmed by the manufacturer, so the consumer had no involvement in its programming (the Manufacturer-Programmed condition). Two additional conditions (which were otherwise identical to the Driver-Programmed condition) included an "override switch" that enabled the driver to intervene, reversing the programmed decision, yielding the Overridden condition (in which the switch was used) and the Override Ignored condition (in which the switch was present but not used).

Thus, the study had a 2 (outcome: selfish or selfless) × 5 (Control System: Manual, Manufacturer-Programmed, Driver-Programmed, Overridden, and Override Ignored) within-subjects design. After reading each vignette, participants rated the driver's blame- or praiseworthiness using a slider anchored at -100 (*most blameworthy*) and 100 (*most praiseworthy*). After evaluating all vignettes, participants were debriefed.

Results

Of the 360 surveyed participants, 114 failed just one attention check and an additional 88 failed both attention checks. Participants who failed only one attention check were retained for the reported analyses, yielding a final sample of 246 participants (59.3% female, $M_{\text{age}} = 39.02$). Note that all reported analyses yielded identical patterns of significant effects when tests were conducted with no exclusions ($N = 360$) and strict exclusions ($N = 158$). All measures and conditions for the variables of interest have been reported. Full descriptions of the sampling approach, exclusion decisions, and these tests are contained in Online Supplemental Material; the study design and hypotheses were preregistered (available at http://aspredicted.org/blind.php/?x=h7et2r).

To assess whether overall differences in attribution of responsibility existed among the 10 vignettes, a 2 (outcome) × 5 (control system) repeated-measures analysis of variance was conducted. There were significant main effects of outcome, $F(1, 245) = 980.69$, $p < .001$, partial $\eta^2 = .80$, and control system, $F(4, 245) = 16.59$, $p < .001$, partial $\eta^2 = .06$, and a significant Outcome × Control System interaction, $F(4, 245) = 159.22$, $p < .001$, partial $\eta^2 = .39$. For all reported means in the following hypotheses, negative numbers denote blame and

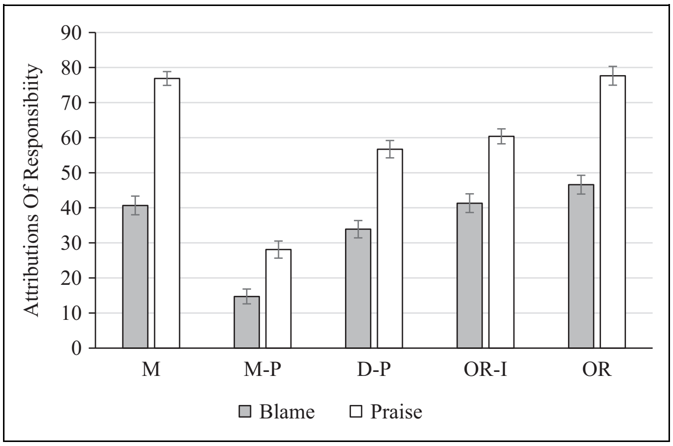

Figure 1. Attributions of responsibility by condition. Error bars depict standard errors. M = manual; M-P = manufacturer-programmed; D-P = driver-programmed; OR-I = override ignored; OR = overridden.

positive numbers denote praise. Figure 1 depicts attributions for all conditions (with blame multiplied by -1, so that attributions could be presented on the same scale). This analysis, however, does not directly test the hypotheses, which pertain to specific differences between vignettes; these were examined using a series of 2 × 2 analyses of variance (ANOVAs) and paired-samples t tests.

To assess the first set of hypotheses, the comparison between manually controlled and manufacturer-programmed vehicles, a 2 (outcome) × 2 (control system) repeated-measures ANOVA was conducted.[1] There were significant main effects of outcome, $F(1, 245) = 739.41$, $p < .001$, partial $\eta^2 = .75$, and control system, $F(1, 245) = 46.59$, $p < .001$, partial $\eta^2 = .16$, and, consistent with predictions, a significant Outcome × Control System interaction, $F(1, 245) = 320.05$, $p < .001$, partial $\eta^2 = .57$. Paired-samples t tests were then conducted to test differences among specific conditions (these paired-samples t tests were conducted after each 2x2 ANOVA). Participants blamed drivers significantly more for selfish outcomes when the vehicle was controlled manually ($M = -40.67$, 95% CI = ± 5.22) than when it was preprogrammed by the manufacturer ($M = -14.71 \pm 4.16$), $t(245) = 9.96$, $p < .001$, Cohen's $d = 0.65$. Similarly, participants praised drivers significantly more for selfless outcomes when the vehicle was controlled manually ($M = 76.88 \pm 3.88$) than when it was preprogrammed by the manufacturer ($M = 28.09 \pm 4.78$), $t(245) = 17.76$, $p < .001$, Cohen's $d = 1.14$.

To assess the second set of hypotheses, the comparison between manufacturer-programmed and driver-programmed systems, a 2 (outcome) × 2 (control system) repeated-measures ANOVA was conducted. There were significant main effects of outcome, $F(1, 245) = 418.88$, $p < .001$, partial $\eta^2 = .63$, and control system, $F(1, 245) = 8.86$, $p = .003$, partial $\eta^2 = .04$, and, as predicted, there was a significant Outcome × Control System interaction, $F(1, 245) = 159.98$, $p < .001$, partial $\eta^2 = .40$. Participants blamed drivers significantly less for selfish outcomes when the vehicle was manufacturer-programmed ($M = -14.71 \pm 4.16$) than when it was driver-programmed ($M = -33.90 \pm 4.83$), $t(245) = 8.56$, $p < .001$, Cohen's $d = 0.55$. Analogously, participants praised drivers significantly less for selfless outcomes when the vehicle was manufacturer programmed ($M = 28.09 \pm 4.78$) than when it was driver programmed ($M = 56.71 \pm 4.84$), $t(245) = 10.70$, $p < .001$, Cohen's $d = 0.68$.

An exploratory hypothesis, the comparison between manually controlled and driver-programmed cars, was also assessed by conducting a 2 (outcome) × 2 (control system) repeated-measures ANOVA. There were significant main effects of outcome, $F(1, 245) = 947.77$, $p < .001$, partial $\eta^2 = .80$, and control system, $F(1, 245) = 22.93$, $p < .001$, partial $\eta^2 = .09$, as well as a significant Outcome × Control System interaction, $F(1, 245) = 62.34$, $p < .001$, partial $\eta^2 = .20$. Participants blamed drivers significantly more for selfish outcomes when the vehicle was manually controlled ($M = -40.67 \pm 5.22$) than when it was driver programmed ($M = -33.90 \pm 4.83$), $t(245) = 2.99$, $p = .003$, Cohen's $d = 0.19$. Correspondingly, participants praised drivers significantly more for selfless outcomes when the vehicle was manually controlled ($M = 76.88 \pm 3.88$) than when it was driver programmed ($M = 56.71 \pm 4.84$), $t(245) = 9.40$, $p < .001$, Cohen's $d = 0.61$.

To examine the third set of hypotheses, comparing driver-programmed vehicles without an override switch to driver-programmed vehicles with an override switch that was ignored, another 2 (outcome) × 2 (control system) repeated-measures ANOVA was conducted. There was a significant main effect of outcome, $F(1, 245) = 683.80$, $p < .001$, partial $\eta^2 = .74$, but not of control system, $F(1, 245) = 1.58$, $p = .21$. As predicted, the Outcome × Control System interaction was significant, $F(1, 245) = 12.17$, $p = .001$, partial $\eta^2 = .05$. Participants blamed drivers significantly more for selfish outcomes when the driver had access to an override switch and did not use it ($M = -41.32 \pm 5.26$) than when the vehicle was driver-programmed and no override switch existed ($M = -33.90 \pm 4.83$), $t(245) = 3.54$, $p < .001$, Cohen's $d = 0.23$. The corresponding difference for selfless outcomes was in the hypothesized direction but was not statistically significant: Participants praised drivers only slightly more for selfless outcomes when they ignored their override switch ($M = 60.39 \pm 5.27$) than when the switch did not exist at all ($M = 56.71 \pm 4.84$), $t(245) = 8.56$, $p = .105$, Cohen's $d = 0.10$.

To assess the final set of hypotheses, comparing driver-programmed vehicles without an override switch to driver-programmed vehicles in which an override switch was used, a final 2 (outcome) × 2 (control system) repeated-measures ANOVA was conducted. There were significant main effects of outcome, $F(1, 245) = 1003.29$, $p < .001$, partial $\eta^2 = .80$, and control system, $F(1, 245) = 7.65$, $p < .01$, partial $\eta^2 = .03$, and as predicted, a significant Outcome × Control System interaction, $F(1, 245) = 84.21$, $p < .001$, partial $\eta^2 = .26$. Participants blamed drivers significantly more for selfish outcomes when the driver initially programmed the vehicle to be selfless but overrode this decision at the critical moment ($M = -46.60 \pm 5.21$) than when the vehicle was driver-programmed selfishly from the outset ($M = -33.90 \pm 4.83$), $t(245) = 5.58$, $p < .001$, Cohen's $d = 0.36$. Analogously, participants praised drivers significantly more for selfless outcomes when they overrode their initial selfish decision ($M = 77.66 \pm 4.18$) than when the vehicle was driver-programmed selflessly from the outset ($M = 56.71 \pm 4.84$), $t(245) = 8.56$, $p < .001$, Cohen's $d = 0.55$.

Discussion

Currently, much is unknown about how the use of technology is viewed from a moral perspective. As technological development continues, investigations in this arena are increasingly important, as humans interact with new technology in ever-more consequential ways. As existing research does not yet provide a strong basis for inferring how moral judgments will play out in the context of new technology, it is important to test extant theories of morality to determine whether they accurately capture (or need revising to accommodate) judgments in this new domain. In addition, experimental manipulations of agency often use cases that are unusual, on the margins of everyday experience. Conversely, AVs by their nature entail a reduction of agency and thus provide a pragmatic context in which to examine its impact. In particular, AVs can reduce agency directly (i.e., the owner literally has no control) or indirectly (i.e., the owner exerts some degree of control at the time of purchase but none after). Thus, studying morality within this framework has both theoretical and practical significance.

The current findings are consistent with research showing that agency reduction induces less blame for negative consequences and praise for positive consequences (Gray & Wegner, 2009; Malle, Guglielmo, & Monroe, 2014; Pizarro, Uhlmann, & Bloom, 2003) and that commissions receive stronger attributions than omissions (Spranca, Minsk, & Baron, 1991). It is worth noting that these effects held in a new paradigm in which the judged agents' future moral agency was outsourced. Specifically, when the AV was programmed by the manufacturer, less responsibility was attributed to the driver than when the AV's programming was chosen by the driver or when the car was controlled manually. Drivers of manually controlled vehicles were deemed more responsible than drivers of driver-programmed vehicles, in contrast to research showing that impulsive immoral acts are perceived as less blameworthy than deliberate immoral acts (Pizarro, Uhlmann, & Salovey, 2003). This finding could be due to the driver's perceived inability to change the outcome (Buckwalter & Turri, 2015), or could be specific to agency reduction within the context of AVs. That drivers who chose a selfish algorithm and ignored an override switch were deemed more responsible than if the switch was not present at all provides further support for the importance of the driver's perceived ability to change an outcome.

Finally, judgments of responsibility were consistent with the notion of "reformed sinners" and "lapsed saints" (Harford & Solomon, 1967). Drivers were blamed more after first making a selfless decision and subsequently overturning that decision, compared to making the selfish choice at the outset. Similarly, drivers were praised more after first making a selfish decision and subsequently overturning that decision, compared to making the selfless choice at the outset. To our knowledge, this is the first demonstration of these effects on judgments of moral responsibility. This could result from beliefs that the driver's last choice was particularly diagnostic of their moral character (i.e., their "true self"; Newman, De Freitas, & Knobe, 2013; Siegel, Crockett, & Dolan, 2017), as evidenced by their effort to change the initial choice (Bigman & Tamir, 2016). Future research would benefit from examining potential mechanisms in this context. Additionally, judgments of agents as reformed sinners and lapsed saints may extend beyond this specific paradigm; continued research could lead to insights about the boundary conditions or explanatory mechanisms underlying these effects more generally. This finding also suggests that the study of moral judgments, which often focuses on attributions made about an agent's single choice, should more often examine judgments of a deeper corpus of behavior. For example, it is likely that having more information about the relevant actions that preceded the outcome influences judgments of the agent's overall character. This makes sense if the inferences made from such judgments are related to how the agent might think or behave in the future; such inferences of moral character are important because they communicate whether one's intentions are oriented toward helpful or harmful behavior (Goodwin, 2015). A more person-centered approach may better reflect real-world moral psychology (Pizarro & Tannenbaum, 2011; Uhlmann et al., 2015), as judgments of any single event are likely unable to be isolated from knowledge of the actor's character.

There are several important limitations to the current research. First, the experiment relied on a within-subjects design. A between-subjects design in which participants do not have the opportunity to make explicit contrasts might yield different results[2] (Charness, Gneezy, & Kuhn, 2012). However, as noted, restricting the analysis to the first vignette participants read, creating a de facto between-subjects design, produced similar results. Additionally, although people may indeed encounter such scenarios in the real world as isolated incidents, it is likely that they will make implicit comparisons to salient alternatives (e.g., judging how culpable a "driver" of an AV should be compared to a driver of a manual vehicle). Furthermore, it is not entirely clear whether and how control system features will be regulated; a future world in which multiple control systems exist in parallel suggests the value of a within-subjects design, as people may indeed be making explicit comparisons of culpability in an analogous way. Second, the study did not provide any information about the agents or victims. As previous research has demonstrated, varying cognitive abilities (e.g., changing a target's age) can lead to different assessments of moral agency (Gray & Wegner, 2009), thus changing the criteria for blame and praise. Future research could use this paradigm to understand how people will assign responsibility to those who outsource their moral agency but vary in cognitive ability and social status.

Last, participants were not given the opportunity to blame or praise an agent other than the driver. This may at first seem to be a strange caveat. However, as AV technology advances, more and more people will not be in direct control of their vehicles, and a higher percentage of traffic accidents will be locally "agentless." Although the owner of the vehicle may not have immediate control at the critical moment, depending on the way in which the vehicle was programmed, they may still be judged as if they *did* have control. Thus, in the future, there could be a shift in the public's thinking about how to assign responsibility, especially blame. Will the most blameworthy party be the consumer, lawmakers, manufacturers, or the team of programmers whose code decided who lives and dies? Should all parties share equal responsibility? Current theories do not fully account for multiple agents being perceived as morally responsible for a single event (for a discussion of causal responsibility, see Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013). Thus, it is important to understand, theoretically and pragmatically, when moral responsibility will be allocated among multiple agents and how that process occurs. Although problems of this sort are often discussed in the context of law (Girgis, 2013; Martin & Newhall, 2016), the relevant psychological processes have been less examined. Future research in this area could be especially useful for examining attributions of responsibility, as technological advancements make it increasingly difficult to delineate sufficient conditions for moral agency.

The current research is the first to investigate how attributions of blame and praise are made in the context of AVs. It replicates, strengthens, and generalizes existing theories about judgments of moral responsibility. Additionally, it provides the first direct evidence that perceptions of agents as reformed sinners or lapsed saints could intensify attributions of responsibility. Pragmatically, understanding how features of AV programming change judgments of responsibility implicates perceptions of AVs among consumers, the media, and the judicial system. Such perceptions have consequences in various contexts, ranging from adoption rates to legal culpability. As AVs are currently being engineered and decisions about how to program them will soon be made, the current research provides important initial evidence about how judgments of blame and praise may play out in the context of emerging technology.

Authors Note

Ryan M. McManus developed the study concept. Both authors contributed to the study design. Testing and data collection were performed by Ryan M. McManus. Ryan M. McManus performed the data analysis and interpretation under the supervision of Abraham M. Rutchick. Both authors drafted, revised, and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. In preregistering the study, we specified that the hypotheses would be tested using paired-samples t-tests. However, because of the symmetrical nature of the conditions, the 2 × 2 interactions offered a more complete test of the hypotheses, and so we present these first, before examining the paired-samples t-tests comparing different control systems within the same outcome. Note that all preregistered simple effects were significant except for the comparison between the driver-programmed and override-ignored conditions for selfless outcomes.
2. To address possible order effects in the within-subjects design, we created a de facto between-subjects design by restricting the data to participants' responses to the first vignette they read. Mean differences were in the hypothesized direction for six or seven (depending on the exclusion criterion chosen) of the eight hypothesized comparisons (and one of the two exploratory comparisons). Given the small sample sizes, these data should be interpreted cautiously, but suggest that order effects were not the primary driver of the observed differences.

Supplemental Material

The supplemental material is available in the online version of the article.

References

Baron, J., & Ritov, I. (1994). Reference points and omission bias. *Organizational Behavior and Human Decision Processes*, *59*, 475–498.

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell handbook of judgment and decision making* (pp. 478–515). Chichester, England: Wiley.

Bentham, J. (1983). *The collected works of Jeremy Bentham: Deontology, together with a table of the springs of action; and the article on utilitarianism*. Oxford, England: Oxford University Press. (Original work published 1879)

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*, 1654–1669.

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

Boudette, N. E., & Markoff, J. (2016, July 1). *The fully self-driving car is still years away*. Retrieved from https://www.nytimes.com/2016/07/02/business/international/bmw-tesla-self-driving-car-mobileye-intel.html?_r=0

Buckwalter, W., & Turri, J. (2015). Inability and obligation in moral judgment. *PLoS One*, *10*, 1–20.

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*, 23–43.

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior and Organization*, *81*, 1–8.

Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General*, *145*, 772–787.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Friedrich, J. (2002). On seeing oneself as less self-serving than others: The ultimate self-serving bias? In R. A. Griggs (Eds.), *Handbook for teaching introductory psychology: Vol. 3: With an emphasis on assessment* (pp. 245–247). Mahwah, NJ: Lawrence Erlbaum.

Gao, P., Hensley, R., & Zielke, A. (2014). *A road map to the future for the auto industry*. Retrieved from http://www.mckinsey.com/industries/automotive-and-assembly/our-insights/a-road-map-to-the-future-for-the-auto-industry

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*, 166–171.

Girgis, S. (2013). The mens rea of accomplice liability: Supporting intentions. *Yale Law Journal*, *123*, 460–494.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science*, *24*, 38–44.

Gray, K., & Schein, C. (2017). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*, 1–39.

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*, 1600–1615.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, *23*, 206–215.

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*, 505–520.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364–371.

Harford, T., & Solomon, L. (1967). Reformed sinner and lapsed saint strategies in the prisoner's dilemma game. *Journal of Conflict Resolution*, *XI*, 104–109.

Kant, I. (2002). *Groundwork for the metaphysics of morals*. New Haven, CT: Yale University Press. (Original work published 1797)

Kordes-de Vaal, J. H. (1996). Intention and the omission bias: Omissions perceived as nondecisions. *Acta Psychologica*, *93*, 161–172.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *37*, 1036–1073.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*, 147–186.

Martin, B., & Newhall, J. (2016). Technology and the guilty mind: When do technology providers become criminal accomplices? *The Journal of Criminal Law & Criminology*, *105*, 95–148.

Matyszczyk, C. (2016). *How will driverless cars make life or death choices? Google exec admits he doesn't know*. Retrieved from https://www.cnet.com/news/how-will-driverless-cars-make-life-or-death-choices-google-exec-admits-he-doesnt-know/?curator=TechREDEF

Mill, J. S. (1863). *Utilitarianism*. London, England: Parker, Son, and Bourne.

Millar, J. (2014). *An ethical dilemma: When robot cars must kill*, *who should pick the victim?* Retrieved from http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/

Newman, G. E., De Freitas, J., & Knobe, J. (2013). Beliefs about the true self explain asymmetries based on moral judgment. *Cognitive Science*, *39*, 96–125.

Open Roboethics Initiative. (2014). *If death by autonomous car is unavoidable*, *who should die? Reader poll results*. Retrieved from http://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*, 653–660.

Pizarro, D., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*, 267–272.

Scanlon, T. M. (1998). *What we owe to each other (Vol. 66)*. Cambridge, MA: Belknap Press of Harvard University Press.

Searle, J. (1983). *Intentionality*. Cambridge, MA: Cambridge University Press.

Sedikides, C., & Alicke, M. D. (2012). Self-enhancement and self-protection motives. In R. M. Ryan (Ed.), *The Oxford handbook of human motivation* (pp. 303–322). New York, NY: Oxford University Press.

Siegel, J., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition*, *167*, 201–211.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, *1*, 229–243.

Spranca, M., Minsk, E., & Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, *27*, 76–105.

Swann, W. J., Griffin, J. J., & Predmore, S. C. (1987). The cognitive-affective crossfire: When self-consistency confronts self-enhancement. *Journal of Personality & Social Psychology*, *52*, 881–889.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*, 1249–1254.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81.

Author Biographies

Ryan M. McManus is a graduate student at California State University, Northridge. His research examines how values and moral judgments can be influenced by social context.

Abraham M. Rutchick is an associate professor of psychology at California State University, Northridge. His research examines nonconscious processes and social perception.

Handling Editor: Jesse Graham