Feelings of Culpability:

Just Following Orders versus Making the Decision Oneself

Maayan S. Malter[1], Sonia S. Kim[1], and Janet A. Metcalfe[2]

[1]Marketing Division, Columbia Business School, Columbia University, and [2]Department of Psychology, Columbia University

Manuscript accepted for publication at Psychological Science

Corresponding Author: Maayan S. Malter, Columbia University, Columbia Business School, Marketing Division, 3022 Broadway, New York, NY 10027-6902
E-mail: mmalter22@gsb.columbia.edu

Statement of Relevance

This paper investigates the relation between agency and felt culpability for behavior. Two conditions were contrasted: (1) the individual makes a decision and takes action; (2) the same action is taken in compliance with an authority figure's decision. Although the legal system holds individuals who make the decision more culpable for a negative outcome (per Nuremberg defense), our results indicated that individuals who 'obeyed orders' *felt* more culpable (responsible, guilty, and regretful). This outcome was a reversal of our own expectations and occurred both in a hypothetical programming of a self-driving car scenario, as well as in two moral-dilemma scenarios regarding COVID-19. Better understanding the conditions in which decision agency will increase or decrease feelings of moral culpability has important implications not only for firms developing autonomous systems/products but also for policy and lawmakers who are setting the protocol for how we, as a society, act and judge moral responsibility.

Abstract

In five experiments (N=1,490), participants were asked to imagine themselves as programmers of self-driving cars who had to decide how to program the car in a potential accident: spare the driver or spare pedestrians. Alternatively, participants imagined they were a mayor, grappling with difficult moral dilemmas concerning COVID-19. Either they, themselves, had to decide how to program the car or which COVID-19 policy to implement (high agency), or they were told by their superior how to act (low agency). After learning that a tragic outcome occurred due to their action, participants reported their felt culpability. Although we expected people to feel less culpable about the outcome if they acted in accordance with their superior's injunction than if they made the decision themselves, participants actually felt more culpable when they followed their superior's order than when they made the decision themselves. Some possible reasons for this counterintuitive finding are discussed.

This paper investigates the relation between agency and culpability for behavior enacted at a distance. Previous research has shown that the feeling of agency increases a person’s involvement (Bandura, 1997), improves their memory (Cloutier & Macrae, 2008; Murty, DuBrow, & Davachi, 2019), and produces a feeling of responsibility (Frith, 2014; Haggard & Tsakiris, 2009). Presumably, if there is a link between the sense of agency and feelings of responsibility, an individual should feel responsible for adverse outcomes to the extent that she decides what to do rather than simply following another’s orders. This common understanding—that guilt and credit are both connected to the individual’s sense of agency—lies at the core of the hypothesis investigated here: that high, rather than low, feelings of agency will result in the individual feeling more responsible and culpable should an adverse outcome occur.

To investigate this issue, two kinds of moral dilemmas were studied: one hypothetical and the other all too real. The first dilemma employed an adaptation of the classic Trolley Car Problem (Cushman, Young, & Hauser, 2006; Foot, 1967; Greene, 2016), in which an engineer at a self-driving car company must program the car to make what Amit and Greene (2012) called a deontological (save the rider above all else) or a utilitarian (save as many people as possible even if it means sacrificing the rider) action. Participants were instructed to imagine that they were a programmer who had high agency (they chose how to program the car) or low agency (they executed their boss’s orders). The second dilemma investigated a moral conundrum faced by communities all over the world as a result of the COVID-19 pandemic, for example whether to remain strictly closed in an effort to thwart the spread of the virus, or to partially open in order to protect other aspects of people's health and welfare. In both situations, the focus was on an individual’s feelings of their own culpability when a negative outcome occurred, as a function of whether they, themselves made the decision about what to do versus when they took action based on the injunction of a superior.

Experiment 1: Self-driving Cars

Method

Participants

Two hundred and one participants (111 females; age: $M = 35.89, SD = 11.24$) were recruited to complete an online survey through Amazon's Mechanical Turk (M-Turk: www.mturk.com). We recruited roughly one hundred participants per agency condition based on a previous M-Turk study with a similar experimental design and context (Amit and Greene, 2012). Additionally, using G*Power software (Faul, Erdfelder, Lang, & Buchner, 2007), we computed that the necessary sample size for F-tests with a moderate effect size, $\alpha = .01, (1 - \beta) = 80$, is 191 participants. All experiments were approved by the authors' university Institutional Review Board (AAAD9781) and all materials are reported on the authors' Open Science Framework page: (https://osf.io/c7yrv/?view_only=746d6d083bbe498bb591fa77d0695ae9). For all experiments reported here, no participants who completed the entire study were excluded from the analyses.

Procedure

Participants were asked to imagine that they were an engineer working at a company developing autonomous cars, and that they must program the car's action in a potential accident. The car could be programmed to be deontological (save the rider above all else) or utilitarian (save as many people as possible even if this harms the rider).

Participants were randomly assigned to one of two conditions: high agency vs. low agency. In the *high agency* condition, they read that the CEO of their company told them that it was solely their choice, as the engineer, to program the car to be deontological or utilitarian. They were then asked to decide how they would program the car. In the *low agency* condition, participants read that the CEO had told them (i.e., it was not their own decision to make) how to program the car—to be either deontological or utilitarian (randomized among low agency condition participants).

Next, all participants were told to imagine that five years had passed since they programmed the car and were asked to describe a day in their lives five years in the future. Following this day-in-the-future exercise, participants read that five years had indeed passed and that one of the cars they had programmed was involved in an accident: the car stopped in front of a crosswalk where five pedestrians were crossing, when a truck lost control of its brakes and was about to hit the car from behind. Those who programmed the car to be deontological read "Because you programmed [were told to program] the car to have deontological behaviors, it changed lanes and caused the five pedestrians to die." Those who programmed the car to be utilitarian read "Because you programmed [were told to program] the car to have utilitarian behaviors, it stayed in front of the truck and caused the rider to die."

Participants then responded to two questions, which we will not consider further, concerning their general affect, that is how they felt about their programming and whether they were upset about what happened, followed by the three questions, central to our study, that were directed at their appraisal of their own moral culpability: "To what extent do you feel you are *responsible* for what happened?", "To what extent do you feel *guilty* about what happened?", and "To what extent do you *regret* programming the car the way you did?" Responses were given on

7-point Likert scales. Participants were then asked, “If you could redo the programming decision [make the programming decision yourself], would you program the car differently?” Finally, basic demographic information was collected.

Results

In all experiments presented here, the results from the three items that were directed at participants’ felt culpability were averaged to produce a single aggregate measure. (The individual data for all questions asked, along with the direction of coding, are provided in the supplementary material.) When given a choice (in the high agency condition), 66% chose the utilitarian (vs. deontological) option. Although the decision type (deontological vs. utilitarian) in the high agency condition was determined by the participant, in the analyses that follow, decision type was treated *as if* it were an independent variable. Because cell sizes were unbalanced, Type III sum of squares were used in the ANOVAs.

A 2 (agency: low vs. high) × 2 (decision type: deontological vs. utilitarian) ANOVA, shown in the left panel of Figure 1, indicated that there was a main effect of agency. Participants in the low agency condition, that is, those who did what the CEO directed, felt more morally culpable ($M_{low\ agency} = 4.26, SD = 1.71$) than did those in the high agency condition who made the decision themselves ($M_{high\ agency} = 3.51, SD = 1.68$; $F(1{,}197) = 12.96, p <$ $.001$, η_p^2=.04). There was a main effect of decision type, such that those who had programmed the car to be deontological (vs. utilitarian) felt more morally culpable ($M_{deontological} =$ $4.19, SD = 1.93$ vs. $M_{utilitarian} = 3.66, SD = 1.54$; $F(1{,}197) = 8.04, p = .005$, η_p^2=.01). In addition, there was an interaction between agency and decision type ($F(1{,}197) = 5.26, p =$ $.023$, η_p^2=.02). Scheffe’s post-hoc test indicated that people in the low (vs. high) agency condition felt more culpable particularly when the car was deontological ($p = .006$), but not so much when it was utilitarian ($p = .925$).

A similar pattern was observed when participants were asked if they would change their programming if they could do it again. A logistic regression revealed a main effect of agency ($\beta_{agency} = -1.14, 95\%\ CI\ [-2.12, -0.23], z = -2.38, p = .017, OR = 0.32, 95\%\ CI\ [0.21, 0.80]$); more participants in the low agency condition would have changed. There was a main effect of decision type ($\beta_{decision} = -2.25, 95\%\ CI\ [-3.44, -1.25], z = -4.10, p < .001, OR = 0.10, 95\%\ CI\ [0.03, 0.29]$) such that more participants who made deontological decisions wanted to change. In addition, there was a marginal interaction between agency and decision type ($\beta_{agency \times decision} = 1.45, 95\%\ CI\ [-0.01, 3.00], z = 1.91, p = .057, OR = 4.27, 95\%\ CI[0.99, 20.15]$). When the car was deontological, 52% in the low agency (vs. 26% in the high agency) condition said that they would program the car differently if they had to do it over. When the car was utilitarian, 10% in the low agency (vs. 13% in the high agency) condition indicated that, if given the opportunity, they would have changed the programming.

Insofar as more people in the low agency than the high agency condition indicated that they would have changed their programming, we investigated the possibility that the primary result of interest might have been entirely attributable to people having been forced to go against their own inclinations. The question concerning whether they would have changed provides some indication of what participants would have decided had they been given free choice. This is not a perfect measure, though, since there could have been hindsight-related changes in both groups once the tragic outcome was known. And, indeed, even in the high agency condition several participants said *yes*, they would have programmed the car differently. Accordingly, in a follow-up ANOVA only participants, in both groups, who said *no* they would not have changed the programming were included. The effect of interest, that low agency participants felt more culpable than high agency participants, was still observed ($M_{high\ agency} = 3.17, SD = 1.60; vs. M_{low\ agency} = 3.70, SD = 1.51, F(1{,}148) = 13.93, p = 0.017, \eta_p^2 = .04$). The main effect of decision type and the interaction between decision type and agency were not significant ($p > .10$). Being required to go against one's own inclinations is very likely important, and worthy of further consideration. However, this analysis indicated that this factor alone is not sufficient to explain the primary result of interest, namely that people who obeyed a superior's orders felt more, rather than less, culpable than did people who made the decision themselves.

Experiment 2: Replication and time lag

The effect shown in Experiment 1, in which people who obeyed the orders of their superior felt more culpable for a negative outcome than did people who made the decision themselves, was unexpected. In the literature, immediate outcomes are associated with greater feelings of agency than are delayed outcomes (Metcalfe & Greene, 2007; Michotte, 1946). Perhaps one factor contributing to the counterintuitive result in Experiment 1 was that people felt little agency, even in the high agency condition, because the outcome was delayed. Experiment 2, then, had two aims: (1) to replicate the results of Experiment 1, and (2) to investigate whether the delay or immediacy of the outcome affected felt responsibility.

Method

Four hundred and four participants were recruited through M-Turk (167 females; age: $M = 37.63, SD = 12.26$). Experiment 1 yielded a moderate effect size, so we continued to use the practice of recruiting one hundred participants per condition for Experiments 2-5. The method for Experiment 2 was identical to that of Experiment 1, except for an additional factor, temporal lag. After imagining programming the car, participants were randomized into two conditions: immediate vs. delay. In the *immediate* condition, they read: “Right after you program the self-driving car, it goes directly out onto the city streets with real people (where its actions have real consequences) for testing. As the programmer, you are observing the car through a drone flying overhead, when the car encounters an inevitable accident.” The *delay* condition was identical to Experiment 1, with the accident occurring five years later. An additional open-ended question was given at the end of the survey: “Why do you think you feel this level of regret? Briefly explain your reasons.”

Results

Seventy percent of participants in the high agency condition chose to program the car to execute a utilitarian (vs. deontological) behavior. A 2 (agency: low vs. high) × 2 (decision type: deontological vs. utilitarian) × 2 (temporal lag: immediate vs. delayed) ANOVA indicated, as shown in the right panel of Figure 1, that there was, again, a main effect of agency such that participants in the low agency condition felt more culpable ($M_{low\ agency} = 4.53, SD = 1.71$) than did those in the high agency condition ($M_{high\ agency} = 3.70, SD = 1.74;\ F(1,396) = 9.42\ , p = .002, \eta_p^2 = .06$). There was no effect of temporal lag ($M_{immediate} = 4.33, SD = 1.77$ vs. $M_{delayed} = 3.89, SD = 1.75;\ F(1,396) = 2.09,\ p = .149$) although directionally, participants felt more culpable in the immediate than the delayed condition. The effect of decision type in this experiment was not significant ($M_{deontological} = 4.18, SD = 1.95$ vs. $M_{utilitarian} = 4.07, SD = 1.65; F(1,396) = 1.08\ , p = .298$). There was no interaction between agency and decision type ($F(1,396) = 1.68, p = .195$), although the direction was the same as was found in Experiment 1.

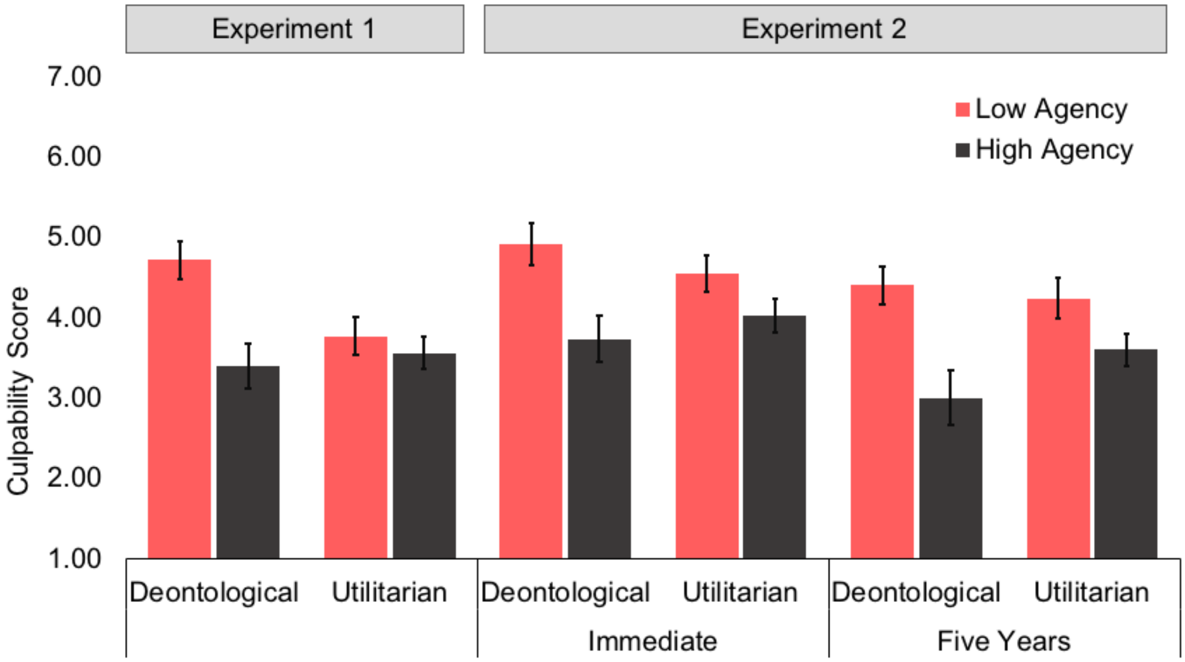

Figure 1. *Results for Self-Driving Car Scenarios in Experiments 1 and 2. The culpability score combines responses for felt responsibility, guilt, and regret. In the deontological condition 5 people died; in the utilitarian condition the rider died. The error bars represent the standard error of the mean in each condition.*

Experiment 3: From the outside perspective

Both Experiments 1 and 2 showed results that were the opposite of our initial predictions. The unexpected results may have occurred because participants, in imagining that they were the programmer, had taken a first-person perspective, whereas in making our predictions, we had taken a third-person perspective. To test whether this vantage-point difference impacted people’s evaluations of responsibility, in Experiment 3 participants were induced to take a third-person perspective.

Method

One hundred and one participants were recruited through M-Turk. No demographics were collected. Participants read a detailed description of Experiment 1, including a description of what Experiment 1 participants read in the low agency condition (referred to as group A) and the high agency condition (referred to as group B). The utilitarian and deontological decision types were described. Participants reported their predictions about how individuals in group A and B would respond to the three questions related to culpability, namely, how much responsibility, guilt, and regret the programmer would feel about the situation. They gave their predictions on the same 7-point Likert scales that participants had used in the previous experiments. Participants were only asked to make predictions about the effect of agency. They were not asked about the effect of decision type or about any interactions.

Results

As shown in the left panel of Figure 2, participants predicted that a programmer would feel more morally culpable in the high agency condition ($M_{high\ agency} = 5.73, SD = 1.30$) than in the low agency condition ($M_{low\ agency} = 3.53, SD = 1.50;\ t(100) = -11.92, p < .001, 95\%\ CI = (-2.58, -1.83);\ d = 2.36, 95\%\ CI = (2.13, 2.60)$). This is the opposite of the results from Experiments 1 and 2, but consistent with our pre-experimental predictions and the hypothesis that the first-person perspective in the first two experiments was important.

Experiment 4: Identification with the Agent

Experiment 4 described the four conditions from the scenario of the preceding experiments but referred to the programmer as "Alex." By personalizing and asking how "Alex" would feel, we thought to increase participants' identification with the programmer. Then, with the same participants, a third-person perspective was induced by asking them for their own judgments about Alex's responsibility.

Method

Four hundred and one participants were recruited through M-Turk (213 females; age: $M = 38.42, SD = 12.28$), and randomly assigned to one of four conditions in a 2 (agency: low vs. high) × 2 (decision type: deontological vs. utilitarian) between-subjects design. Participants read the same scenario described in Experiments 1 and 2, but instead of imagining themselves as the programmer, they read that a programmer named Alex either made the decision to program the car or was told to program the car to be deontological or utilitarian. Participants were first asked to indicate how they thought Alex would respond to the same questions used in Experiments 1 and 2, then participants were asked to judge from their own perspective (as a third-person observer of the situation) Alex's behavior: "To what extent do *you* think Alex is responsible for what happened?"

Results

Participants' evaluations of how culpable Alex would feel are shown in the middle panel of Figure 2. There was no effect of agency ($M_{low\ agency} = 4.76, SD = 1.47$ vs. $M_{high\ agency} = 4.84, SD = 1.60;\ F(1{,}397) = 0.73, p = .394$). There was an effect of decision type ($M_{deontological} = 5.09, SD = 1.47$ vs. $M_{utilitarian} = 4.55, SD = 1.56; F(1{,}397) = 4.45, p = .036$, $\eta_p^2 = .01$), such that participants judged that Alex would feel more culpable when more people were killed. The interaction between agency and decision type was not significant ($p = .569$).

Interestingly, as shown in the right panel of Figure 2, when participants were asked how responsible they, themselves, judged Alex to be, they assessed him as more responsible when he was in the high agency versus the low agency condition ($M_{high\ agency} = 3.89, SD = 1.92$ vs. $M_{\ low\ agency} = 3.34, SD = 1.85; F(1{,}397) = 4.29, p = .039$, $\eta_p^2 = .001$). They also judged Alex as more responsible if he programed the car to execute a deontological compared to utilitarian action ($M_{deontological} = 3.99, SD = 1.90$ vs. $M_{utilitarian} = 3.28, SD = 1.85; F(1,397) = 7.91, p = .005, \eta_p^2 = .009$). There was no interaction ($p = .964$).

To examine how changing one's vantage-point for making moral judgments would influence a participant's evaluation of Alex's responsibility, the dependent measure 'feelings of responsibility' was submitted to a linear mixed effects regression (LMER; 'lmerTest' package in R) with agency and decision type as between-subjects factors, and perspective as a within-subjects factor. There was a main effect of perspective ($\beta_{perspective} = -0.96, 95\%\ CI\ [-1.29, -0.63],\ p < .001$) such that participants thought Alex would feel more responsible when they imagined how he felt (first-person), compared to when *they* were judging Alex's responsibility (third-person). There was also an interaction between perspective and decision type ($\beta_{perspective \times decision} = -0.48, 95\%\ CI\ [-0.94, -0.01], p = .044$). Importantly, excluding decision type from the model revealed a marginally significant interaction between perspective and agency ($\beta_{perspective \times agency} = 0.31, 95\%\ CI\ [-0.02, 0.63],\ p = .067$). Responsibility judgments did not differ between low and high agency conditions when participants were thinking of how Alex would have felt, but were higher in the high (vs. low) agency condition when they made their own judgment from a third-person perspective.

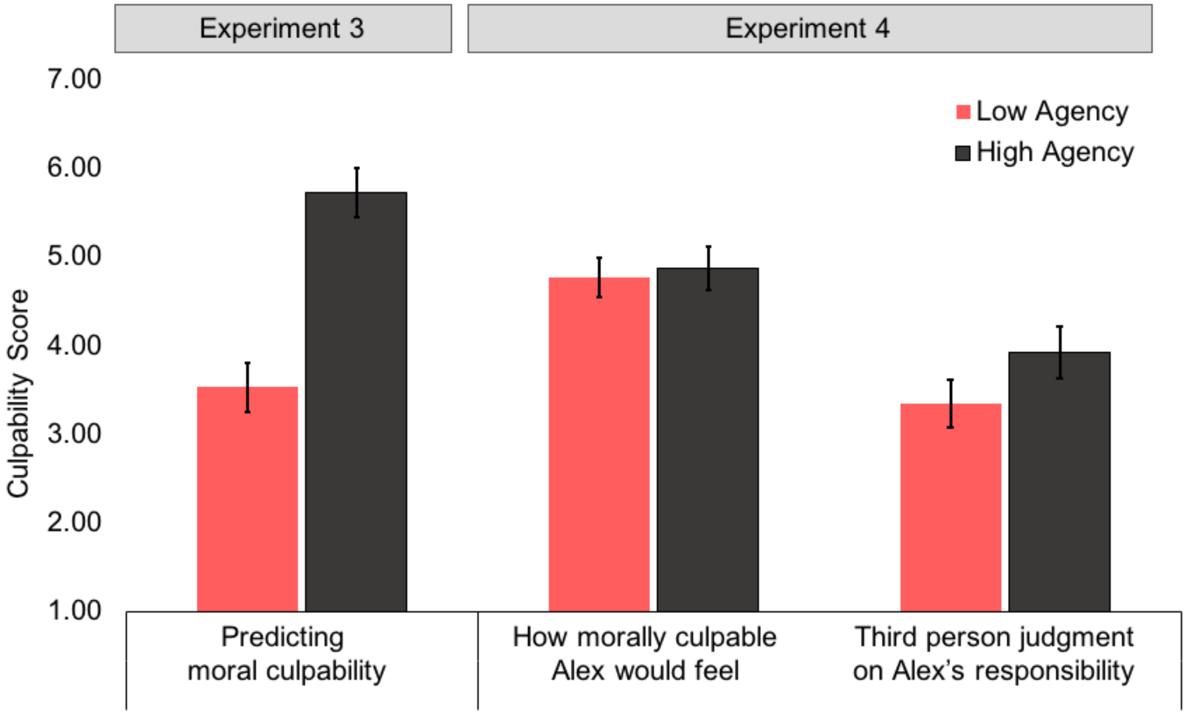

Figure 2. *Left: Results for Experiment 3; participants' predictions of how morally culpable a programmer in Experiment 1 would feel in the high agency vs. low agency condition; Right: Results for Experiment 4; (a) participants' evaluation of how morally culpable Alex would feel and (b) participants' own judgment of Alex's responsibility for the outcome of the situation. The error bars represent the standard error of the mean in each condition.*

Discussion of Experiments 1-4

Results from the first two experiments indicated that when individuals imagined being the person who enacted a moral trade-off, they felt *more* culpable when they had obeyed the orders of someone else than when they made the decision themselves. However, as shown in Experiments 3 and 4, when judging the programmer's actions from the outside, the programmer who made their own decision was viewed to be more morally culpable. This latter finding, indicating that it is the person who makes the decision, rather than the one who simply carries out the orders, who is responsible, is consistent with the superior orders or 'Nuremberg defense' (Green, 1976). First- versus third-person perspective appeared to be crucial to the pattern of results. When asked how ‘Alex’ would feel, the results were mixed (and may have depended upon whether the participant identified with Alex or judged Alex from the outside).

The question remains as to why, in the first-person perspective, people experienced greater feelings of culpability when they had just obeyed orders as compared to when they had made the decision themselves. Interestingly, Frith (2014) noted that there are two factors that impact people’s feelings of responsibility: their feeling of agency, but also the counterfactual thought that they might have acted differently. Participants in the low agency condition may have had a different experience of this second factor than those in the high agency condition who carefully deliberated about their choice. Participants’ open-ended responses, including comments such as “I should have pushed back on the CEO,” suggested such counterfactual thought. The finding in Experiment 1—that more participants in the low agency condition said they would program the car differently if given another chance—suggests that counterfactual thinking likely played some role. Nevertheless, the results were the same when those who said they would have acted differently were excluded. While this factor may have played a role, other factors may have also contributed. Experimentally delineating the reasons for people’s exaggerated feelings of culpability in what are ostensibly passive situations remains a challenge with both theoretical and practical implications. So too is the question of whether such a reversal in feelings of culpability plays out in more realistic situations where people's lives are palpably at stake.

Experiment 5: COVID-19 Policy Decisions

The first four experiments were conducted before the COVID-19 pandemic struck, and before the advent of the wrenching moral dilemmas that, with the pandemic, were brought into widespread and acute focus. Experiment 5 investigates the same issues relating to the effects of agency on feelings of culpability as did Experiments 1 and 2, but this time applied to the kinds of real-world decisions that were being hotly debated and enacted at the time the experiment was conducted, impacting the welfare of real people including the participants themselves.[1]

To better isolate the effects of agency on culpability, the outcome (vis, the number of deaths) in Experiment 5 was the same regardless of the decision. Two moral dilemma scenarios were used to establish greater generality of the findings. Scenario 1 posed the question: Should bars and restaurants reopen for in-person dining? Reopening (lenient option) would help the local economy but risk spreading the virus, whereas remaining closed (strict option) would prevent spreading the virus but negatively impact the local economy. Scenario 2 posed the question: Should delivery workers be required to be tested for COVID-19 in order to deliver to elderly/at-risk consumers? Note that, at the time of the experiment, COVID-19 tests were very scarce. Allowing workers to make deliveries without testing (lenient option) would ensure that consumers received essential items such as food and prescription medicines but risk spreading the virus, whereas requiring testing (strict option) would prevent spreading the virus but risk individuals not receiving necessary supplies.

Method

Participants

Three hundred and eighty-three American participants were recruited through M-Turk (199 females, 179 males, 5 non-binary; age: $M = 32.15, SD = 10.46$).

Procedure

The experiment was a 2 (scenario: 1 or 2) × 2 (agency: high vs. low) × 2 (policy option: strict vs. lenient) between-subjects design. All participants were asked to imagine that they were

---

[1] Distinct neural circuitry has been found to be activated when an individual makes a hypothetical versus real moral decision (FeldmanHall, Dalgleish, Thompson, Evans, Schweizer, & Mobbs, 2012).

a mayor of a mid-sized American city, where they had to make important decisions regarding the pandemic.

Participants read the descriptions (pros and cons) of each policy option in one of the two scenarios (reopening bars and restaurants, or delivery worker testing). Participants were randomly assigned to the high agency or low agency condition. In the *high agency* condition, they read that there were no state-level guidelines and thus it was their role, as the mayor, to choose the strict or lenient option. They were then asked to indicate their choice. In the *low agency* condition, participants, read that the governor issued state-level guidelines for which policy must be implemented and were randomly assigned to either the lenient or the strict option.

Next, all participants read how the policy they implemented played out over the following month. **Scenario 1**: Those who implemented the *strict* policy to keep bars and restaurants closed read that the policy caused the local economy to struggle, resulting in 20 deaths due to people not being able to afford basic necessities, while those who implemented the *lenient* policy to reopen bars and restaurants read that the policy caused the spread of the virus, resulting in 20 deaths. **Scenario 2**: Those who implemented the *strict* policy to require testing for delivery workers read that the policy limited the supply of eligible delivery workers, leading to 20 deaths from lack of food and medicine and other non-COVID-19 causes, while those who implemented the *lenient* policy to allow untested delivery workers read that the policy caused the virus to spread, resulting in 20 deaths.

After reading the entire scenario, participants responded to the same questions as in Experiments 1 and 2. Finally, participants responded to the COVID-19-threat scale (Kachanoff et al. 2020), and reported their political affiliation (both on a continuous scale: very-liberal to very-conservative, and party affiliation) along with demographic information.

Results

Fifty-four percent of participants in the high agency condition (56% and 51%, respectively, for scenarios 1 and 2) chose the strict option (vs. lenient option). As shown in Figure 3, there was, again, a main effect of agency such that individuals in the low agency condition felt more culpable ($M_{low\ agency} = 4.65, SD = 1.42$) than did those in the high agency condition ($M_{high\ agency} = 4.09, SD = 1.47$; $F(1{,}375) = 3.69$, $p = .057$, $\eta_p^2 = .009$). There was no effect of scenario ($F(1{,}375) = 1.06$, $p = .303$, $\eta_p^2 = .003$). There was an effect of policy option ($M_{strict} = 4.08, SD = 1.40$ vs. $M_{lenient} = 4.68, SD = 1.48$; $F(1{,}375) = 9.19$, $p = .003$, $\eta_p^2 = .022$), such that people in the lenient condition felt more culpable. There was no interaction between agency and policy option ($p = .787$).

A model including a continuous measure of political affiliation was tested. There was no main effect of political affiliation nor was there an interaction between political affiliation and policy option ($ps > .1$). There was an interaction between agency and political orientation ($F(1{,}373) = 11.88$, $p < .001$, $\eta_p^2 = .03$), such that the effect of agency was larger for more liberal individuals. The three-way interaction between policy option, political affiliation and agency was not significant ($p = .325$). Because this was not the focus of the present study and because the reasons for the interaction and its replicability are not yet known, it will not be discussed further at this time.

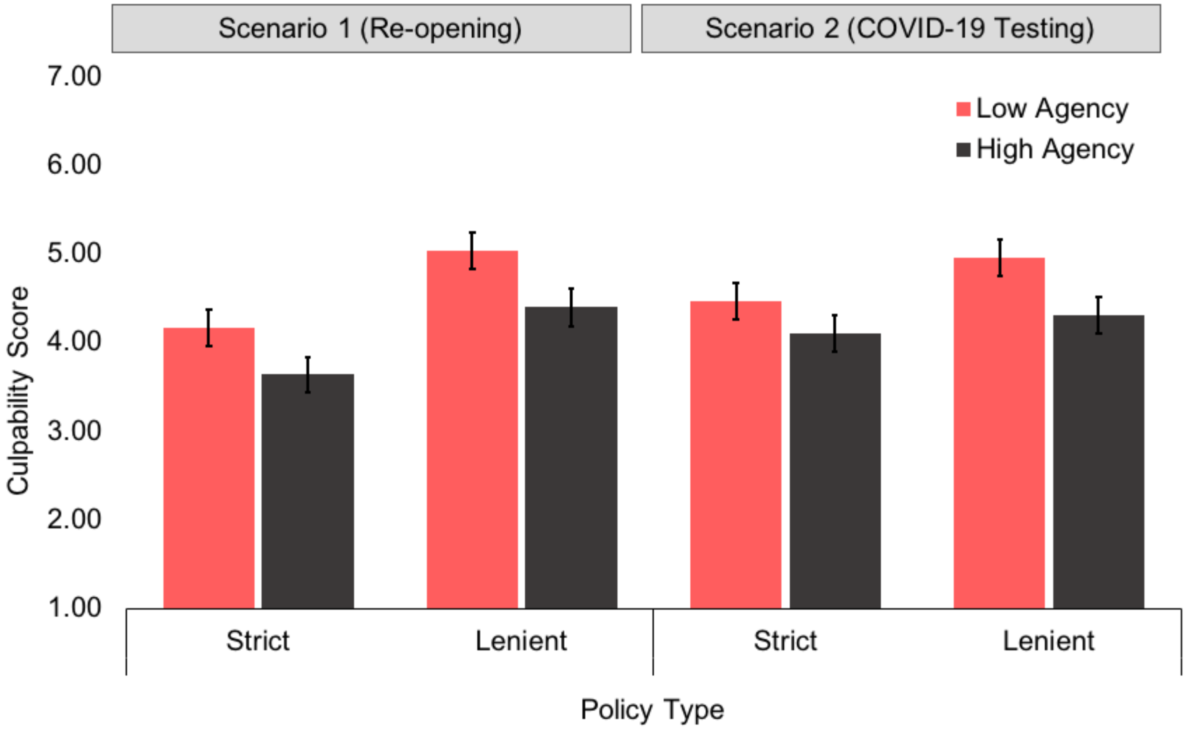

Figure 3. *Results for COVID-19 Scenarios in Experiment 5, which investigated a real-world situation in which a mayor made her own policy decision (high agency) or was told which policy to implement by the governor (low agency) about reopening the economy after the initial pandemic shutdown. The error bars represent the standard error of the mean in each condition.*

Summary of Effects

A summary of culpability effect sizes across all five experiments is presented in figure 4, comparing first-person and third-person perspectives. A positive value indicates that high agency showed higher culpability than low agency, and a negative value indicates that low agency showed higher culpability than high agency, as given by Cohen's D. Conditions coded as being from a first-person perspective included moral culpability responses from Experiments 1, 2, and 5, and people's ratings of how Alex (the programmer) would feel in Experiment 4 (total of 1,389 first-person data points). In classifying the feelings of Alex as first person, we made the assumption that people imagined themselves *to be* Alex, although, as noted in the introduction and discussion of Experiment 4, this condition was an ambiguous case. Data that were coded as being from a third-person perspective included participants' predictions about how culpable other people would feel (from Experiment 3), and participants' evaluations of the extent to which they judged Alex to be responsible for the outcome, in Experiment 4 (total of 603 third-person data points). The effects in the first-person data compared to the third-person data are in the opposite directions from zero and the error bars are non-overlapping. These results indicate that from a third-person perspective, high agency individuals were deemed to be more culpable than were low agency individuals, whereas from the first-person perspective the reverse occurred: low agency participants felt themselves to be more culpable than did high agency participants.

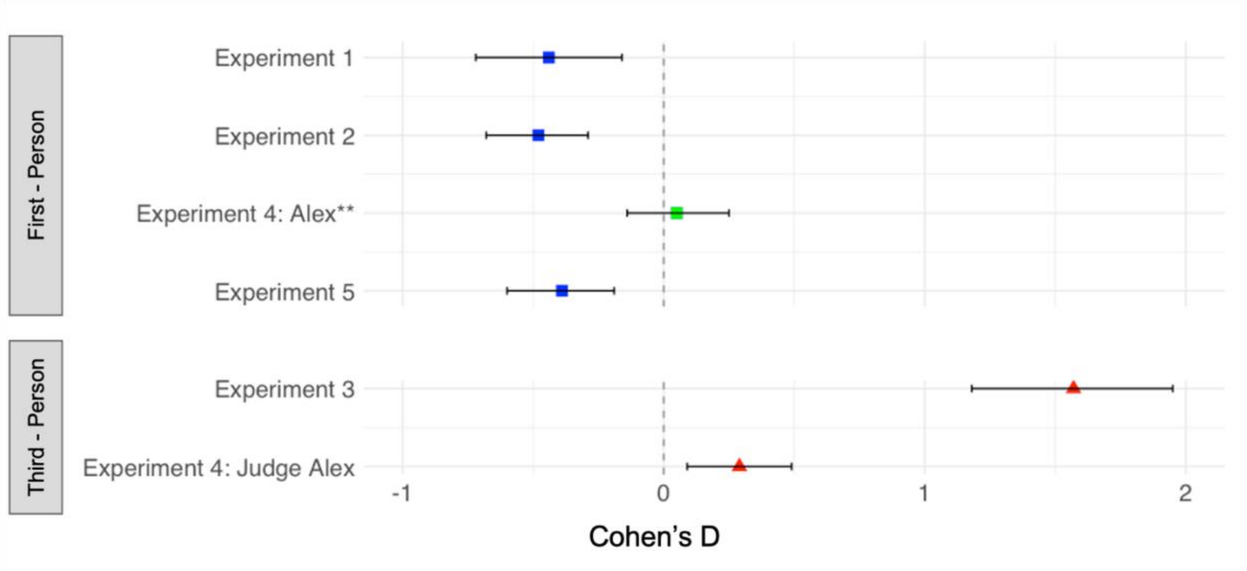

Figure 4. *Forest plot with the effect sizes (Cohen's D) of the effect of agency on moral culpability across studies with different perspectives. For data points with a third-person perspective, agency has a positive effect on culpability. For data points with a first-person perspective, agency has a negative effect on culpability (**for Experiment 4, please see discussion in text). Error bars represent the 95% confidence intervals of each Cohen's D value.*

Conclusion

The idea that action driven by one's own intention is more deserving of praise, if the outcome is favorable, and blame, if it is not, than is non-intentional action, is fundamental to both our legal system and our personal attributions of moral credit and culpability. As far back as the 13th century BC, Hittite law inscribed in cuneiform on stone tablets asserted this distinction concerning the intent of the doer: "If anyone blinds a free man in a quarrel, he shall give one mina of silver. If only his hand does wrong he shall give 20 shekels."[2] Even preschool children make distinctions based on intent (Carpenter, Akhtar, Tomasello, 1998; Cushman, Sheketoff, Wharton, Carey, 2013), and the construct is foundational to many psychological theories of moral attribution (e.g., Alicke, 2000; Bonicalzi, & Haggard, 2019; Bucknoff, 2020; Caspar, Christiensen, Cleeremens, & Haggard, 2016; Cushman, 2008; Malle, Guglielmo & Monroe, 2014; Piaget, 1932; Weiner, 1995). Although the legal system is generally concerned with ill intent (*mens rea*), credit is also differentially accorded for favorable outcomes following actions that are intentional (i.e., when one makes a choice) as compared to those in which one only carries out another's decision.

The data from Experiments 3 and 4 are consistent with this intention or agency-based attribution of responsibility. When the programmer of the self-driving car was judged from the outside, the programmer was deemed less culpable for a tragic outcome when the decision was made by their superior (the CEO of the company), than when they made it themselves. The person who acted at the behest of another is viewed, both in the law and in Experiments 3 and 4, as less culpable than the one who actually made the decision.

---

[2] One mina of silver was worth sixty shekels. Interestingly, this distinction was apparently not recognized three centuries earlier. Article V from the 16th century BC tablet read only: "If anyone blinds a free man or knocks out his teeth he shall pay 20 shekels of silver." (Tablets and translations from the Anadolu Medeniyetleri Muzesi, in Ankara, Turkey).

The surprise in the results presented here, in Experiments 1, 2 and 5, is that when people imagined themselves to be the person who made the decision as compared to implementing the will of their superior, they felt the opposite: they did not feel less culpable after having obeyed the orders of another, they felt more culpable. Perhaps this result occurred because they imagined that if they had been given free choice they would have chosen differently or that the outcome might have been different and less tragic. There are hints in the data that suggest that they sometimes felt that they should have pushed back. Perhaps they felt that they had violated their own moral standards and would not have made the same decision about the act that they had carried out. They may have felt conflict about their behavior. Perhaps they just felt that they should have given more consideration to an action with such grave consequences. However, even among participants who indicated that they would not change the decision that they implemented, people who followed orders still felt more culpable than those who made the decision themselves. Whatever the reason, their feelings of guilt, responsibility and regret were amplified, not muted, by acquiescing in another's decision. Although outsiders blamed them less, they blamed themselves more.

These five experiments show that “just obeying orders” does not get one off the hook to *oneself*. In fact, the opposite was observed: individuals felt more culpable for negative outcomes when they had low, rather than high, agency in making a decision with moral tradeoffs. These findings, while counterintuitive, appear to be robust. They replicated in a hypothetical trolley-car situation adapted to programming a self-driving car, and generalized to two up-to-date and real-life COVID-19 situations. Importantly, the relation between agency and attribution of moral culpability appears to depend on whether the individual views the actions from a first-person or third-person perspective. While from a third-person perspective, people are held to be less culpable when they carry out the decision of another, it is striking that from a first-person perspective they, nevertheless, *feel more culpable*. Understanding the feelings that people experience after either making decisions themselves or following another's orders in distressing moral situations has important consequences not only for policy makers, but also for the mental health and well-being of the individuals who carry out such decisions.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556-574.

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science*, *23*(8), 861-868.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Macmillan.

Bonicalzi, S., & Haggard, P. (2019). Responsibility between neuroscience and criminal law: The control component of criminal liability. *Rivista Internazionale di Filosofia psicologia, 10,* 103-119.

Bucknoff, Z. (2020) Motivator and moralizer: How agency shapes choices and judgement. [Unpublished doctoral dissertation] Columbia University.

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen-through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development*, *21*(2), 315-330.

Caspar, E. A., Christensen, J. F., Cleeremans, A., & Haggard, P. (2016). Coercion changes the sense of agency in the human brain. *Current Biology*, *26*(5), 585-592.

Cloutier, J., & Macrae, C. N. (2008). The feeling of choosing: Self-involvement and the cognitive status of things past. *Consciousness and Cognition*, *17*(1), 125-135.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353-380.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6-21.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological science*, *17*(12), 1082-1089.

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.

FeldmanHall, O., Dalgleish, T., Thompson, R., Evans, D., Schweizer, S., & Mobbs, D. (2012). Differential neural circuitry and self-interest in real vs hypothetical moral decisions. *Social Cognitive and Affective Neuroscience*, *7*(7), 743-751.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review*, 5, 5–15.

Frith, C. D. (2014). Action, agency and responsibility. *Neuropsychologia, 55*, 137-142.

Green, L.C. (1976), *Superior Orders in National and International Law*, Netherlands: A. W. Sijthoff International Publishing Co.

Greene, J. (2016). Solving the trolley problem. *A companion to experimental philosophy*, 175-178.

Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feelings, judgments, and responsibility. *Current Directions in Psychological Science*, *18*(4), 242-246.

Hodson, G., MacInnis, C. C., & Choma, B. L. (2019). Left-right differences in perspective-taking across US states. *Personality and Individual Differences*, *144*, 36-39.

Kachanoff, F. J., Bigman, Y. E., Kapsaskis, K., and Gray, K. (in press), "Measuring Realistic and Symbolic Threats of COVID-19 and Their Unique Impacts on Well-Being and Adherence to Public Health Behaviors," *Social Psychological and Personality Science*.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186.

Metcalfe, J., & Greene, M. J. (2007). Metacognition of agency. *Journal of Experimental Psychology: General*, *136*(2), 184.

Michotte, A. (1946). La perception de la causalité. Paris: Vrin.

Morris, S. G. (2020). Empathy and the Liberal-Conservative Political Divide in the US. Journal of Social and Political Psychology, 8(1), 8-24.

Murty, V. P., DuBrow, S., & Davachi, L. (2019). Decision-making increases episodic memory via postencoding consolidation. *Journal of Cognitive Neuroscience*, *31*(9), 1308-1317.

Piaget, J. (1932/1965). *The Moral Development of the Child*. New York, NY: Free Press.

Sparkman, D. J., & Eidelman, S. (2016). “Putting myself in their shoes”: Ethnic perspective taking explains liberal–conservative differences in prejudice and stereotyping. *Personality and Individual Differences*, *98*, 1-5.

Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. New York: Guilford.