Citation:
Gantman, A P., Sternisko, A., Gollwitzer, P. M., Oettingen, G,. Van Bavel, J. J. (*in press).* Allocating moral responsibility to multiple agents. *Journal of Experimental Social Psychology.*

Allocating Moral Responsibility to Multiple Agents

Ana P. Gantman

New York University

Anni Sternisko

New York University

Peter M. Gollwitzer

New York University and University of Konstanz

Gabriele Oettingen

New York University and University of Hamburg

Jay J. Van Bavel

New York University

Word Count: 6,744

Acknowledgments: For help with manuscript preparation, data analysis, and writing we would like to thank Nirupika Sharma. For critical manuscript feedback we thank Dave Gantman, and members of the Gantman Lab and the Social Identity & Morality Lab for their thoughtful comments on earlier drafts of this manuscript. For help with data collection and study design we would like to thank Justin Lieberknecht. This research was partially supported by a National Science Foundation Grant (#1349089) to JJVB.

Author Contributions: All authors contributed to research design; APG and AS performed research; APG and AS analyzed data; APG, AS and JJVB wrote paper with critical edits from PMG and GO.

Abstract

Moral and immoral actions often involve multiple individuals who play different roles in bringing about the outcome. For example, one agent may deliberate and decide what to do while another may plan and implement that decision. We suggest that the Mindset Theory of Action Phases provides a useful lens through which to understand these cases and the implications that these different roles, which correspond to different mindsets, have for judgments of moral responsibility. In Experiment 1, participants learned about a disastrous oil spill in which one company made decisions about a faulty oil rig, and another installed that rig. Participants judged the company who made decisions as more responsible than the company who implemented them. In Experiment 2 and a direct replication, we tested whether people judge implementers to be morally responsible at all. We examined a known asymmetry in blame and praise. Moral agents received blame for actions that resulted in a bad outcome but not praise for the same action that resulted in a good outcome. We found this asymmetry for deciders but not implementers, an indication that implementers were judged through a moral lens to a lesser extent than deciders. Implications for allocating moral responsibility across multiple agents are discussed.

Word count: 203/250 words

Keywords: morality, intentionality, side effect effect, action phases, causality

Allocating moral responsibility to multiple agents

On April 20, 2010, a Deepwater Horizon oil rig exploded in the Gulf of Mexico, causing the most detrimental oil spill in American history. Although owned by Transocean, the rig was leased by oil company BP, who failed to initiate the rig's fail-safe. The question of culpability quickly arose. One report emphasized that BP made final decisions regarding installation and should shoulder the blame, despite Halliburton having installed the malfunctioning rig (Pallardy, 2018). BP acted as the decider, presumably weighing pros and cons, and making all final decisions regarding the oil rig. Halliburton was the implementer, carrying out the decisions that came from BP. Ultimately, a U.S. district judge allocated more of the blame to BP (67%) than Halliburton (30%). Blame and punishment primarily fell to those who deliberated on and decided what to do, and secondarily to those who planned and implemented those decisions. In this paper, we examine how people allocate moral responsibility when there are multiple potentially culpable agents who play different roles. We have two primary research aims: First, we test whether people hold the decider more morally responsible than the implementer. Second, we investigate whether people judge implementers to be moral agents to a lesser extent than deciders.

Action Phases and Mental States

According to the Mindset Theory of Action Phases, a single action can be divided into discrete phases before and after a goal is set (Gollwitzer, 2012). While choosing which goal to pursue, a person's mindset is deliberative, weighing pros and cons as they decide what to do. Once a goal has been selected, the person's mindset changes to an implemental one, facilitating planning and reaching the goal (Heckhausen, 1986; Gollwitzer, 1999; Gollwitzer, Heckhausen, & Steller, 1990). Here, we suggest that these action phases can be spread across multiple agents. In the above example, BP was responsible for deliberating about and *deciding* on what to do;

Halliburton was in charge of planning and *implementing* the decisions made by BP. We suggest that the different roles played by the two companies led to differential legal punishment, with BP shouldering more of the responsibility than Halliburton.

We present the Mindset Theory of Action Phases as a powerful analogy for understanding multi-agent moral responsibility. The Mindset Theory of Action Phases—when applied to multiple agents—predicts that greater responsibility will be allocated to the decider as the deliberative mindset ends with the setting of an intention. Second, the Mindset Theory of Action Phases, when applied to multiple agents, raises an important question: How do people allocate moral responsibility for implementers? We examined this issue in a series of experiments.

The decider is held morally responsible

The Mindset Theory of Action Phases makes predictions about moral responsibility for the decider that are consistent with contemporary theories of responsibility, especially when a negative outcome occurred (Alicke, 2000; Cushman, 2008; Malle, Guglielmo, & Monroe, 2014). The decider 's role is to weigh pros and cons and, importantly, to make a final decision about what to do (i.e., sets the intention that initiates goal striving). As such, the decider meets classic criteria for moral responsibility and punishment because the decider has culpable mental states, specifically relevant beliefs, desires, and intentions (for a review, see Cushman, 2008; Alicke 2000).

Why might the implementer be held morally responsible?

We suggest that there are three primary reasons why people may also allocate some responsibility to the implementer. First, implementers have relevant mental states regarding the outcome they brought about. Implementers think about where, when, and how to achieve the

goal specified by the decider. Implementers know what they are trying to accomplish (if not why). Second, implementers are direct causes of the outcome of the action and causal connection is important for judgments of responsibility (i.e., blame and punishment; but not wrongness judgments; Cushman, 2008). For example, when an agent had another agent carry out their harmful intent for them, they were assigned *less* moral responsibility than when they went ahead and carried out the harmful action themselves (Paharia, Kassam, Greene, & Bazerman, 2009). Therefore, implementers may also be judged to be deserving of blame and praise

Third, people seem to allocate praise to those who implement. Implementers not only carry out the actions of others with negative consequences, but positive ones too. People strategically blame one but praise many (Schein, Jackson, & Gray, 2019). In one study, people read a vignette in which multiple agents were involved in a risky investment. When the investment failed, people concentrated moral responsibility on the CEO, who ultimately made the investment decision. When the investment was successful, people allocated moral responsibility more evenly and considered agents in consulting and executing roles praiseworthy. In other words, the deliberating agents received all the blame but shared moral praise with others involved in the action. This suggests that people do see the relevant causal role that implementers play, though perhaps they reserve their judgments of moral responsibility for negative outcomes for deciders alone.

Why might the implementer* not *be morally responsible?

There is also evidence to suggest that implementers may *not* be held morally responsible. Specifically, implementers do not necessarily consider whether their actions are right or wrong, they are judged relatively less responsible when the negative outcome was intended by the decider, and they may not be judged to be essential for bringing about the outcome or

representative of the larger group of people who brought about the outcome. We discuss each of these in turn.

People may view implementers as causally responsible, but not morally responsible. Even though implementers have mental states *about* the action, they may not have mental states about whether the action is right or wrong. Indeed, they may never consider it, as the implemental mindset blinds agents to reasons to quit, making them more resilient in pursuing their goals (Gollwitzer, 2012). Accordingly, if an implementer is causally involved in bringing about a bad outcome, they may be judged as deserving of punishment for that outcome, but not necessarily as having done something morally wrong because they did not possess the relevant (i.e., culpable) mental states for moral condemnation (Cushman, 2008). It is even possible that without the relevant mental states for culpability, implementers could be viewed as mechanical causes of the outcome rather than as teleological ones. While we usually consider people's actions *teleologically*, meaning in terms of *why* they did what they did, implementers may be judged *mechanistically*, in terms of *how* they brought about the outcome (Lombrozo, 2010). If so, they will likely not be considered responsible for the outcome in a moral sense.

Moreover, in other research regarding multi-agent responsibility, some (usually the powerful) can manifest their intentions through the actions of others via manipulation. In one set of studies, participants read about a situation in which the government of a small nation caused a food shortage that led workers to attack a small village. When the government intentionally caused the food shortage to manipulate the workers, participants judged the government to be more responsible than the workers. When the government did not intentionally cause the shortage and manipulate the workers, then the workers were assigned more blame (Phillips &

Shaw, 2015). When the government acted intentionally—as deciders do—the responsibility of workers, who implemented the intent of the government, was reduced.

It is also possible that implementers will not be held morally responsible for the outcomes they bring about because implementers may not be judged as pivotal to the outcome. This means that if implementers had behaved differently, the outcome would have been the same (Zultan, Gersenberg, & Lagnado, 2012). For instance, more pivotal team members were assigned greater blame for the loss of the game than team members whose performance was less pivotal to the outcome (Zultan, Gerstenberg, & Lagnado, 2012). It is likely that people do not consider implementers to be pivotal because they are viewed as interchangeable—a decider can often find another agent to implement their intentions if the first one refuses. Thus, while the *role* of the implementer is critical for the outcome to have occurred, the individual who implemented it may be viewed as interchangeable with others and therefore seen as less pivotal to the outcome.

Finally, implementers may not be viewed as representative of the larger agent who made the decision. Previous research has found that for positive outcomes, people judge an agent who is more representative of the company (i.e., the president vs. a low-level clerk) to be more praiseworthy for a good outcome even if they did not explicitly cause it (Zemba & Young, 2012).

Current Research

The current research sought to use the Mindset Theory of Action Phases (Gollwitzer, 2012) to understand attributions of moral responsibility for implementers as well as deciders. We predict that (1) the decider will garner more moral responsibility than the implementer, and (2) that this is because the implementer is regarded as a moral agent to a lesser extent than the decider. We test these hypotheses in two different scenarios that both pertain to affecting the

environment. In Experiment 1, we directly asked people to judge the moral responsibility of all the agents described in the Deepwater Water Horizon Oil Spill. In Experiments 2a and 2b, we examined whether people regard the implementer as less of a moral agent than the decider. To test this question, we capitalized on a known asymmetry in blame and praise among deciders, known as the Side Effect Effect (for a review see Knobe 2010). Deciders get more blame for a negative outcome than praise for the equivalent positive outcome, which has been interpreted as evidence of the primacy of the moral judgments of the deciding agent (Knobe, 2010). Accordingly, in Experiments 2a and 2b, we expected responsibility judgments of the decider, but not the implementer, to be sensitive to whether a positive or negative outcome occurred. As far as we know, this research is the first to measure responsibility for both the decider and the implementer in cases where the Side Effect Effect occurs.

Experiment 1

In Experiment 1, we sought to test whether participants would make distinctions between deciders and implementers when making judgments of moral responsibility for a negative outcome. Accordingly, we presented participants with a vignette very similar to the introductory paragraph of this paper. We described an oil spill that caused deaths, severe environmental damage, and high costs in damages. We described three companies who were all involved in the oil spill: One made the decisions about installing the faulty oil rig, another implemented the faulty oil rig that exploded, and a third owned the rig. This scenario mirrored the real companies involved in the Deepwater Horizon Oil Spill.

Methods

Participants. To determine sample size, we conducted an *a priori* power analysis in G*Power for a repeated measures ANOVA with a within-between interaction effect. From

previous research, we used f = .15 for 80% power, and chose a conservative r = .2 for our repeated measures, which yielded a sample of 117 to achieve 80% power. Given a rough estimate that ~10% of mturkers may not pay attention (and we were willing to pay for extra subs),we upped this number to 130. Prior to data analysis, we noticed incomplete data and opted to up the sample again, resulting in a final 156 participants from Amazon's Mechanical Turk who each received $1.00 for participation. Participants with missing data remained in the dataset unless they did not provide responses for any of the four key dependent measurements ($n_{excluded}$ = 12). The final sample included 144 participants (n_{men} = 80, n_{women} = 55, $n_{trans/GNC}$ = 1 $n_{missing}$ = 8, M_{age} = 34.24, SD_{age} = 10.75). Experiment 1 was conducted in Fall 2018. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012)

We also conducted a *post-hoc* Sensitivity Power Analysis using the software program G*Power 3.1. We set the alpha significance parameter to .05 and the power parameter to .80. We then specified our recruited sample size (N = 156) as well as the observed correlation among the responsibility attributions (average r = .17). This resulted in a critical population effect size of Cohen's f =.13.

Materials and design.

Scenario. Participants read:

> An oil rig explodes in the middle of the ocean, causing one of the most detrimental oil spills in history. The rig that exploded was owned and operated by drilling company **Atlantix** and installed by **Severton.** The rig had a weak concrete core and exploded once natural gas traveled through, killing 15 workers and injuring 26. Although owned by **Atlantix**, the rig was leased by oil company **PetroCorps**,

who failed to initiate the rig's fail-safe designed to close the channel and prevent spills. As a result, millions of barrels of oil leaked into the sea and created a spill spanning four countries, causing billions of dollars in damage and contaminated waters, killing countless animals, and leading to massive unemployment. In this scenario, the question of culpability would quickly arise.

Atlantix owned the rig; **Severton** installed the malfunctioning rig; and **PetroCorps** made all decisions regarding installation.

Design. We created six versions of the scenario that were unique in terms of which company name was assigned to each role but otherwise identical. Participants were randomly assigned to one of the versions.

Attribution of responsibility. We measured judgments of blameworthiness for each company by asking "How much is [company] to blame for the effect on the environment?" on a scale of 0 "Not at all to blame" to 100 "Entirely to blame." We measured judgments of intent by asking "How much did [company] intend to affect the environment?" from 0 "Not at all intentional" to 100 "Entirely intentional." We measured punishment, by asking, "imagine the company incurs some fine from harming the environment. How much should [Company] pay?" on a scale from 0% of [Company's] monthly profit to 100% of [Company's] monthly profit.

Allocation of fixed amount of blame. Finally, we also asked participants to allocate 100% of the blame among the three companies. Participants entered numeric amounts for each company and could not move on from the page until they added up to 100. We included this question because we had access to real data about how much blame was assigned to each

company by a court of law and we wanted to see how the moral blame ascriptions of our participants mirror real legal decision making. See supplemental materials and OSF for full survey.

Procedure. Participants read a brief description of the Deepwater Horizon oil spill in 2010 with the original names changed. Then participants rated each company (in random order) in terms of how much blame they deserved for the spill's effect on the environment, how much they intended to affect the environment, and what percent of their profits they should pay as fine. Then participants were asked to allocate 100% of the blame across the three companies. Finally, participants were asked whether the scenario reminded them of real events (and if so, which one) and responded to a final attention check and demographic questions. All materials and data are available on the Open Science Framework upon publication or upon request by the Editor and Reviewers. (https://osf.io/dsncj/?view_only=e851079c1e524b48addd9d87e66c8a7d).

Results

Participants' responses did not vary as a function of which scenario version they were assigned to, so we collapsed data across the six versions.

Allocation of responsibility. Our three measures of responsibility were blame, intent, and punishment. These three items did not cohere into a single scale item for the decider (⟨ = .53), though they did for the implementer (⟨ = .74) and for the owner of the rig (⟨ =.86). To avoid averaging across variables that are not tapping into the same construct we analyzed each one separately (See the General Discussion for why we think this might have happened).

Blame and punishment. **Patterns for** blame and punishment were very similar and so we report them together here. As hypothesized, the company that made decisions (M_{blame} = 78.81, SD_{blame} = 18.65; $M_{punishment}$ = 74.92, $SD_{punishment}$ = 23.21) was perceived as more

blameworthy and deserving of more punishment than the company who installed the rig (M_{blame} = 60.80, SD_{blame} = 27.78; $M_{punishment}$ = 61.16, $SD_{punishment}$ = 29.90) and, than the company who owned it (M_{blame} = 46.49, SD_{blame} = 30.63; $M_{punishment}$ = 49.18 , $SD_{punishment}$ = 32.46; $F_{blame}(2, 145) = 64.50$, $p < .001$, $\eta^2 = .31$; $F_{punishment}(2, 145) = 40.13$, $p < .001$. $\eta^2 = .22$). Post-hoc comparisons revealed that all groups were different from each other for both blame and punishment, all *p*s < .001.

Intent. A similar overall pattern of results was obtained for judgments of intent, except that participants did not regard the implementer and the owner differently. As hypothesized, the company that made decisions (M_{intent} = 51.15, SD_{intent} = 32.33) was judged as having greater intent to harm the environment than the company who installed the rig (M_{intent} = 38.12, SD_{intent} = 31.95) and the company who owned it (M_{intent} = 35.03, SD_{intent} = 33.18); ($F(2, 145) = 27.36$, $p < .001$, $\eta^2 = .16$). Post-hoc comparisons revealed that participants regard the decider as having more intentionally harmed the environment than both the implementer and the owner of the rig, *ps* < .001. Ascriptions of intent of the implementing company and the owning company did not differ significantly, $p = .21$. [1]

[1] We pre-registered that we would re-run all of our analyses excluding those who failed the attention check, and report on all participants if this did not change the results.

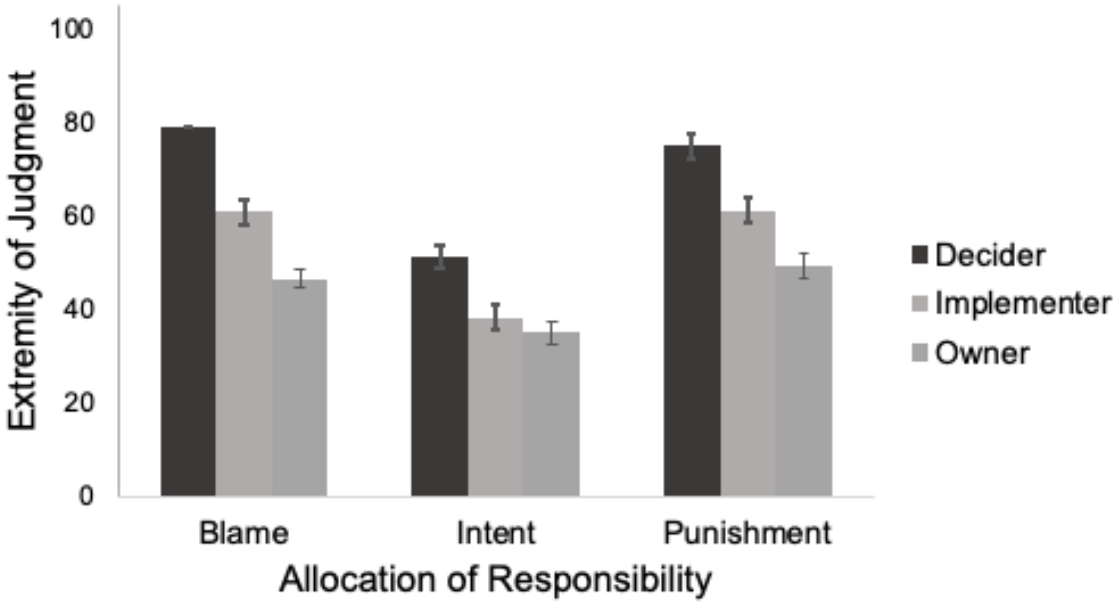

Figure 1. Judgments of intent, blame, and punishment for the company that made decisions about the oil rig (Decider), installed the oil rig (Implementer), and owned the oil rig (Owner), that resulted in a disastrous oil spill. Error bars represent +/- 1 standard error of the mean

Allocation of fixed amount of blame. Overall, participants and the court had a similar pattern of ratings, assigning the most responsibility to the deciding company compared to the company who installed the rig and the company who owned it. Of the possible 100% of blame to allocate, our participants assigned 44% blame to the decider while the court assigned the decider 67% of the blame. Participants and the court system allocated a similar percent of blame for the implementing company (32% vs. 30%) and participants assigned more blame to the owner than did the court system (22% vs. 3%). We present these findings out of interest, with the caveat that our participants and the court system had access to different information, and it is also possible that participants knew about this court ruling and it informed their judgments.

Discussion

We found support for the primary hypothesis. The company who made decisions that brought about a negative outcome was deemed more blameworthy and more punishable and judged as behaving more intentionally than the company who implemented those decisions. Specifically, when asked to think about an oil spill with far-ranging severe consequences, participants held the company that made the decisions about the installation of the oil rig (akin to real life BP) more morally responsible than the company that installed the rig (akin to Halliburton) and the company that owned the rig (akin to Transocean). Like the real court ruling regarding the Deepwater Horizon Oil Spill, participants' attributed more blame to the deciding company (BP) than to the implementer (Halliburton) and the owning company (Transocean). In sum, we found evidence for our first hypothesis: The decider is judged to be more morally responsible than the implementer. Next, we sought to test whether the implementer is judged to be a moral agent to a lesser extent than the decider.

Experiment 2a

In Experiments 2a and 2b, we investigated our second prediction: the implementer is regarded as a moral agent to a lesser extent than the decider. To test this second question, we exploited an existing asymmetry between judgments of blame and praise known as the Side Effect Effect (for a review, Knobe, 2010). The Side Effect Effect occurs when people hold others accountable for morally bad, but not morally good side effects. For example, participants learn that the environment is either harmed or helped by a decision made by the Chairman of the Board of a company. Participants blame the Chairman of the Board when the environment is harmed but offer no praise when the environment is helped. The Side Effect Effect is typically interpreted as evidence of the primacy of moral judgments; we see an asymmetry in judgments

of blame and praise because the Chairman of the Board is being judged through a moral lens (Knobe, 2010).

We modified the original Side Effect Effect (Knobe, 2003) materials to test this question. According to Knobe (2010), the Side Effect Effect results from the primacy of moral judgment. We reasoned that we should see a a greater asymmetry between moral responsibility for good and bad outcomes only for agents who are being judged through a moral lens. We hypothesized that the decider, but not the implementer, would be subject to this asymmetry in blame and praise. We used the original Side Effect Effect vignette and made minor changes to its wording that emphasized the different roles (deliberating and deciding vs. planning and implementing) of the two agents. For the first time in research pertaining to the Side Effect Effect, we asked participants to evaluate the moral responsibility of both the Chairman of the Board, the decider, and the Vice President, the implementer.

Methods

Participants. We recruited 65 participants via Amazon's Mechanical Turk who each received $.25 for participation. The sample size was determined for economic reasons. Participants with missing data remained in the dataset unless they did not provide responses for any of the four key dependent measurements ($n_{\text{excluded}} = 9$). The final sample included 56 participants ($n_{\text{men}} = 31$, $n_{\text{women}} = 25$, $M_{\text{age}} = 31.09$, $SD_{\text{age}} = 10.11$). Experiment 2a was conducted in Fall 2013.

We conducted a *post-hoc* Sensitivity Power Analysis using the software program G*Power 3.1. We set the alpha significance parameter to .05 and the power parameter to .80. We then specified our recruited sample size ($N = 56$) as well as the observed correlation among the

responsibility attributions (r = -.02). This resulted in a critical population effect size of Cohen's f =.27.

Procedure. Participants read one of two modified Side Effect Effect vignettes. They then judged the responsibility of each agent involved. They responded to two questions that measured the responsibility attributed to the Chairman of the Board and two questions that measured the responsibility attributed to the Vice President in randomized order, followed by an attention check and demographics.

Materials and design.

Scenario. We added to Knobe's (2003) original vignettes. Specifically, we added a description of the Chairman of the Board as "deliberating and deciding" and the Vice President as "planning and implementing." The two scenarios were identical except that in one the side effect was morally bad (i.e., the environment was harmed) and in the other, the side effect was morally good (i.e., the environment was helped). Full vignettes are in the appendix.

Attribution of responsibility. The perceived blame-/praiseworthiness and intent were measured for both the Chairman of the Board (the deliberator and decider) and the Vice president (the planner and implementer) and averaged to create a perceived responsibility score. Items taken from the original work: (1) "How much blame/praise does the Chairman of the Board [Vice President] deserve for the effect on the environment?" measured on a scale from 0 (none of the praise/blame) to 100 (all of the praise/blame); (2) "How much did the Chairman of the Board [Vice President] intend to affect the environment?" measured on a scale from 0 (not at all intentional) to 100 (entirely intentional). Accordingly, we arrived at a 2 (Side Effect: harm, help; between-subject) × 2 (Agent: Chairman of the Board, Vice President; within-subject) mixed design.

Results and Discussion

Attributions of blame/praise and intent were highly correlated (for both the Chairman of the Board ($r(54) = .76$, $p < .001$, $\langle = .86$) and the Vice President ($r(52) = .56$, $p < .001$, $\langle = .71$), we computed an average of these ratings for our responsibility score for each agent. We conducted a 2 (Side Effect: harm, help) × 2 (Agent: Chairman of the Board, Vice President) mixed-model ANOVA predicting responsibility attributions. We specified Side Effect (harm vs. help) as between-subjects factor and Agent (Chairman of the Board vs. Vice President) as within-subjects factor.

To examine whether responsibility allocations differ for the two agents, we tested the interaction effect between side effect and agent. We found a significant Side Effect × Agent interaction effect ($F(1,54) = 12.08$, $p = .001$, $\eta^2 = .18$). To de-compose this interaction we examined responsibility attributions separately for the two agents. We first replicated the original Side Effect Effect. Participants attributed more responsibility to the Chairman of the Board when the environment was harmed ($M = 78.02$, $SD = 20.81$) than when it was helped ($M = 43.79$, $SD = 32.77$) ($t(54) = 4.67$, $SE = 7.34$, $p < .001$, CI_{95} [19.52, 48.94], Cohen's $d = 1.25$). We did not find the same pattern for the Vice President. Instead, participants held the Vice President equally accountable when the environment was harmed ($M = 76.77$, $SD = 20.44$) than when it was helped ($M = 75.23$, $SD = 21.03$) ($t(54) = .28$, $SE = 5.54$, $p = .78$, CI_{95} [-9.57, 12.65], Cohen's $d = .07$; see Figure 2).

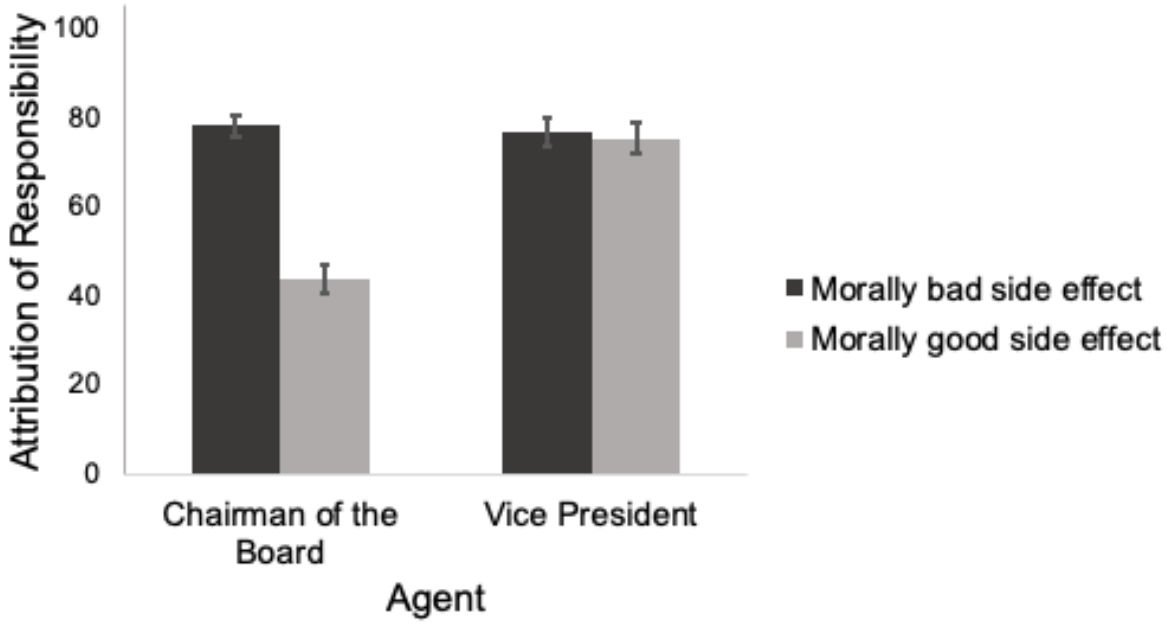

Figure 2. Responsibility attributions for the Chairman of the Board (i.e., the decider) and the Vice President (i.e., the implementer) for morally bad and morally good outcomes. *Significant interaction. Error bars represent +- 1 standard error of the mean.

In Experiment 2a, we found a greater asymmetry in blame and praise for the decider than for the implementer. This suggests that the decider is regarded as a moral agent, but the implementer is not. Notably, the implementer is granted high levels of responsibility in both cases, suggesting that participants recognize their involvement in the outcome (potentially since they are causally linked to the outcome) but they are judged to be responsible in a *moral* sense to a much lesser extent than the decider.

Experiment 2b

Experiment 2a was conducted with a small sample, so we conducted a replication for Experiment 2b with some minor improvements to our methods. In Experiment 2b, we modified the original vignette further to ensure that the roles of both agents are either exclusively

deliberating and deciding (for the Chairman of the Board) or exclusively planning and implementing (for the Vice President). This was important because in the original vignette, the Vice President brings the idea that will help or harm the environment to the Chairman of the Board. It is possible to infer that the plan is the Vice President's idea, and would suggest he had also potentially deliberated about and decided on what to do. In our modification, it is clear that the initiative that will help or harm the environment was the Chairman of the Board's idea, and the decision to move forward with it was his and his alone. In addition, we added a dependent variable that measured desire for punishment or reward. We hypothesized that we would replicate the patterns of Experiment 2a for all three dependent variables. Experiment 2b was conducted in 2018 and pre-registered on OSF (https://osf.io/dsncj/?view_only=e851079c1e524b48addd9d87e66c8a7d).

Methods

Participants. We recruited 160 participants using Amazon's Mechanical Turk. We used G*Power 3.1. to *a priori* determine the required sample size to achieve 80% power. We specified the anticipated effect size as small (Cohen's $f = .15$) and a correlation of $r = .20$ for our repeated measures. This yielded a final sample of 142. Using Amazon's Mechanical Turk to recruit participants, we anticipated ~10% attrition (Simmons, Nelson, & Simonsohn, 2011, 2013) and set the sample size to 160. Four participants were not registered by the recruiting platform but still accessed and completed the survey. The overall sample size was therefore 164 ($n_{men} = 106$, $n_{women} = 58$, $M_{age} = 34.24$[2], $SD_{age} = 11.34$).[3] Experiment 2b was conducted in Fall 2018.

[2] One participant reported to be 2 years old and was excluded from the calculation of the mean age.

[3] For easy comparison to Experiment 2a, we also report the sensitivity power analysis. For attribution of responsibility, we set the alpha significance parameter to .05 and the power parameter to .80. We then specified the recruited sample size (N = 164) as well as the correlation among the responsibility attributions ($r = .05$). This resulted in a critical population effect size of Cohen's $f = .15$).

Material and design.

Scenario. To emphasize the purely implemental role of the Vice President we further modified Knobe's (2003) original vignette. We included the Vice President saying: "As you requested, I looked into the new program you came up with and it will help us increase profits."

Attribution of responsibility. The perceived blame/praise and intent were measured separately for both agents in random order using the same items as in Experiment 2a.

Punishment and reward. Participants were asked about punishment/reward for both agents in random order. They were asked "Imagine that the company incurs some fine/profit from harming/helping the environment. How much should the [agent] pay/get?" and indicated their punishment/reward decision on a scale from 0% of the [agent's] monthly income to 100% of the [agent's] monthly income.

Results and discussion

For each agent, intent and blame/praise ratings were highly correlated and averaged for an overall responsibility score ($r(162) = .80$, $p < .001$, $\langle = .88$ for the Chairman of the Board and $r(162) = .66$, $p < .001$, $\langle = .81$ for the Vice President). Like Experiment 2a, we conducted a 2 (Side Effect: harm, help; between-subject) × 2 (Agent: Chairman of the Board, Vice President; within-subject) mixed-model measures ANOVA predicting responsibility attributions. We specified Side Effect as between-subjects factor and Agent as within-subjects factor. As predicted, we replicated the Side Effect × Agent interaction effect from Experiment 2a ($F(1,162) = 28.56$, $p < .001$, $\eta^2 = .15$).

Responsibility. We decomposed this interaction and examined attribution of responsibility separately for each agent. As in Experiment 2a, we replicated the original Side Effect Effect for the Chairman of the Board such that he received more responsibility attribution

when the environment was harmed (M = 73.16, SD = 23.62) than when the environment was helped (M = 29.02, SD = 29.49, t(147.46) = 10.51, SE = 4.15, p < .001, CI_{95} [35.84, 52.45], d = 1.65). In our more highly powered experiment, the Vice President was also subject to the Side Effect Effect and received more responsibility when the environment was harmed (M = 63.87, SD = 28.20) than when the environment was helped (M = 54.12, SD = 30.83, t(162) = 2.12, SE = 4.61, p = .04, CI_{95} [.65, 18.86], d = .33)[4]. However, the effect size of the asymmetry in attribution of responsibility was much greater for the Chairman of the Board (d = 1.65) than for the Vice President (d = .33). Results are displayed in Figure 3.

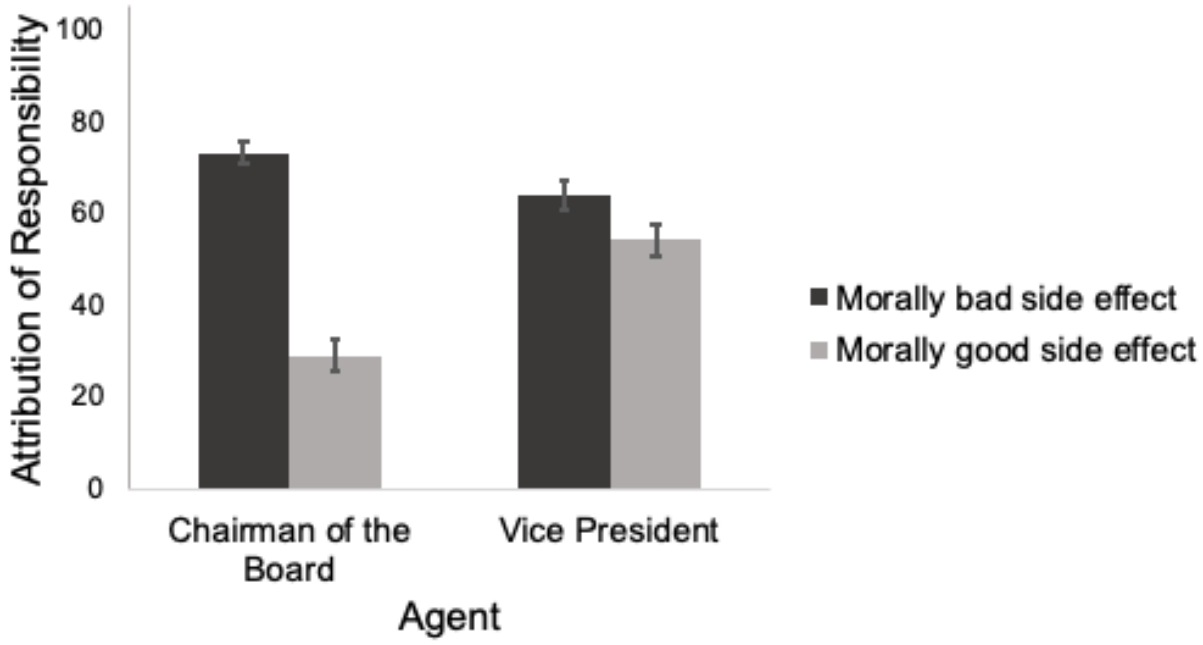

Figure 3. Responsibility attributions for the Chairman of the Board (i.e., the decider) and the Vice President (i.e., the implementer) for morally bad and morally good outcomes. Error bars represent +/- 1 standard error of the mean.

Punishment and reward. We conducted the same analysis as above using punishment/reward as our dependent measure. We observed a significant Side Effect × Agent

[4] We re-ran analyses after excluding participants who failed the manipulation and/or attention check and obtained similar results.

interaction effect ($F(1,162) = 25.37$, $p < .001$, $\eta^2 = .14$). Consistent with results from Experiment 2a, we observed a Side Effect Effect for the Chairman of the Board, such that he was punished more for harming the environment ($M = 65.80$, $SD = 29.98$) than he was rewarded for helping the environment ($M = 29.82$, $SD = 29.88$, $t(162) = 7.69$, $p < .001$, $SE = 4.68$, CI_{95} [26.74, 45.22], $d = 1.20$). Conversely, we did not observe such a Side Effect Effect for the Vice President. There was no significant difference between the magnitude of punishment when the environment was harmed ($M = 52.08$, $SD = 34.06$) and the magnitude of reward when the environment was helped ($M = 46.64$, $SD = 30.91$, $t(162) = 1.07$, $p = .29$, $SE = 5.10$, CI_{95} [-4.63, 15.51], $d = .17$)[5]. Results are displayed in Figure 4.

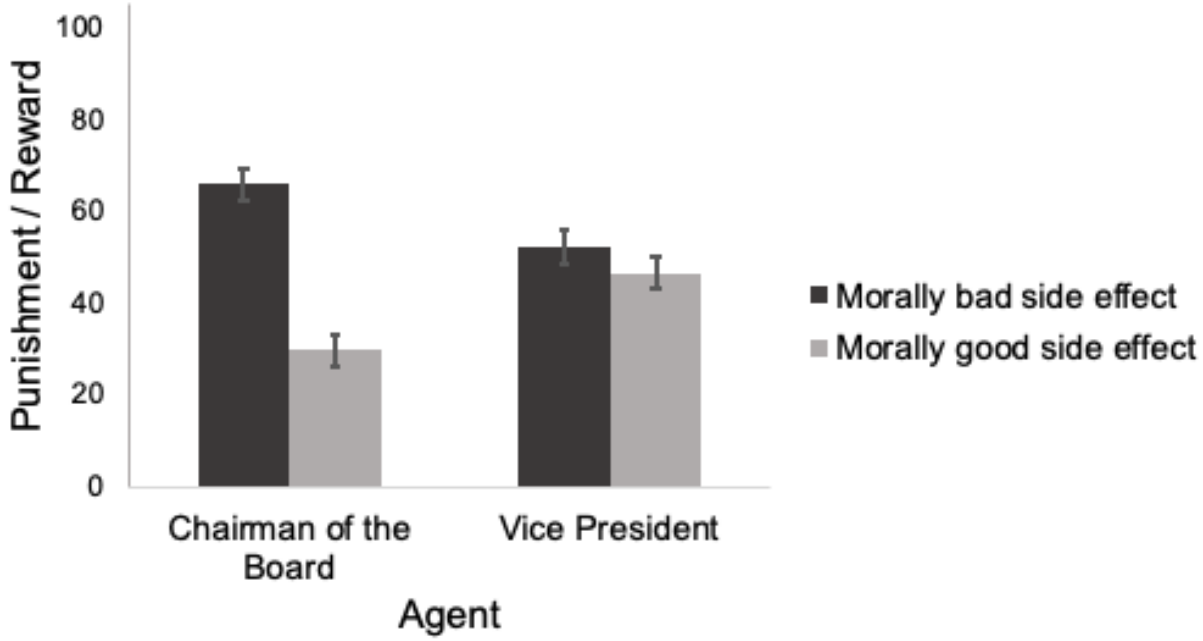

Figure 4. Punishment/reward allocation for the Chairman of the Board (i.e., the decider) and the Vice President (i.e., the implementer) for morally bad and morally good outcomes. Error bars represent +- 1 standard error of the mean

In Experiment 2b, we largely replicated the results of Experiment 2a with a few differences. In Experient 2a we found an asymmetry in blame and praise (i.e., the Side Effect

[5] We re-ran analyses after excluding participants who failed the manipulation and/or attention check and obtained similar results. Notably, the Side Side Effect for the VP was now only marginally significant ($p = .05$).

Effect) for only the decider. In Experiment 2b we found evidence of the Side Effect Effect for both the decider and the implementer, however, the effect was larger for the decider than the implementer. Moreover, only the decider was punished more for a morally bad side effect than he was rewarded for a morally good one (i.e., the environment was helped). These results suggest that the implementer is regarded as a moral agent to a lesser extent than the decider.

General Discussion

Many outcomes involve the work of multiple agents. In this paper, we examined how people allocate moral responsibility to different agents when they play different roles in bringing about the outcome. We proposed that the Mindset Theory of Action Phases (Gollwitzer, 2012) provides a powerful lens for understanding multi-agent responsibility. The mindsets that characterize phases of goal pursuit, especially deciding and implementing, are sometimes divided among agents, and that the roles that correspond to these phases have significant implications for judgments of moral responsibility. We found support for our two main hypotheses: The decider received more moral responsibility and punishment than the implementer for a morally bad outcome (Experiment 1), and the implementer is not regarded as a moral agent to the same extent as the decider (Experiment 2a and 2b). This work advances our understanding of moral responsibility in situations where multiple agents are responsible for the outcome.

In this paper, we used the Mindset Theory of Action Phases to better understand the allocation of moral responsibility when multiple agents are potentially responsible. While many theoretical perspectives converge to predict that responsibility will be largely allocated to deciders, Mindset Theory uniquely raises the question of how to allocate moral responsibility to implementers. Implementers primarily think about where, when, and how to bring about a

desired goal. There is no question whether they are causally involved in bringing about the outcome, but interestingly, because their relevant mental states pertain to *how* to bring about the outcome rather than *whether* it is right or wrong, they are not regarded as moral agents to the same extent as deciders.

Limitations

The current research is silent with regard to the role of power in allocating moral responsibility across multiple agents. Decision-making is often correlated with power, which in turn, confers responsibility. In the oil spill example, the relevant companies do not obviously differ in terms of power in a general sense. However, it seems likely from the scenario that the deciding company paid the implementing one, and so the role of power cannot be ruled out. That said, we do think it is possible for power and decision-making to be decoupled. Indeed, senior aides have been described as the brains behind powerful decision-makers (e.g., political advisor Karl Rove was frequently described as US President George Bush's brain; see Moore & Slater, 2003), and executives often lay the blame for decisions at the hands of their subordinates (e.g., former Enron Chairman blamed his underlings for the scandal that destroyed his company; Goodwyn, 2006). Future work would do well to examine the role of power in these allocations of responsibility.

In all the experiments in this paper, the morally bad outcome was an accident or a side effect rather than the intended outcome. Moral judgments are very sensitive to intent (e.g., Ames & Fiske, 2013; Chalik, Bavel, & Rhodes, 2013; Gray, Young, & Waytz, 2012; Malle et al., 2014; Young & Phillips, 2011). By focusing on side effects in this paper, we leave open the possibility that the implementer was not completely aware of the potential consequences of their actions, and that this could explain why they receive less blame than the decider in Experiment 1. It is

perhaps this feature of our research that resulted in a relatively low internal reliability for judgments of blame, punishment, and intent for the decider in Experiment 1.

Future Directions

We have just begun to understand how people judge the moral responsibility of implementers. We propose two avenues for further research into understanding the relatively small degree of moral responsibility assigned to implementers. First, it is possible that differing causal explanations for the two agents explain differences in attributions of moral responsibility. Specifically, people may explain the behavior of the decider with a teleological "why" explanation, while explaining the behavior of the implementer with a mechanical "how" explanation (Lombrozo, 2010). Given that mechanical explanations tend to apply to objects and machines more than people, this would help us understand why implementers are not viewed as moral agents to the same extent as deciders: their mental states are not salient. Second, we also speculated that the individual who implements may not be viewed as pivotal for the outcome. People may think that the decider would find someone else to carry out their intent if the called upon implementer refuses. It is interesting that the role of implementer--but not the individual who implements—may be pivotal to the outcome. These are areas worthy of more research.

Conclusion

The question of how we allocate moral responsibility across multiple agents is both normatively and descriptively interesting. In the case of the Deepwater Horizon oil spill, the court's judgment matched those of our participants: The decision-maker incurs the greatest share of the blame. Yet, it is important not to mistake these as normative ratings for how we *should* ascribe blame to deciders and implementers. At the trial of Adolf Eichmann, one of the major organizers of the Nazi atrocities in the Holocaust, philosopher Hannah Arendt noted that

Eichmann was not an evil mastermind, but a bureaucrat, someone who thoughtlessly followed another person's orders. It is this observation that led her to coin the phrase "the banality of evil" (Arendt, 1963). Eichmann was an implementer, and Arendt observed that the law's necessity to ascribe him full intent in order to convict him to the fullest degree was failing to capture what had happened.Our findings here, highlight why her observations were met with so much backlash (Ezra, 2007). She was misunderstood as wanting to blame Eichmann less for his crimes, but she was really challenging our intuitions that implementers are not morally responsible. By extending the Mindset Theory of Action Phases across multiple agents, we were able to find evidence for that very intuition.

References

BP's reckless conduct caused Deepwater Horizon oil spill, judge rules. (2014). *The Guardian.* Retrieved from: https://www.theguardian.com/environment/2014/sep/04/bp-reckless conduct-oil-spill-judge-rules.

Ezra, M. The Eichmann Polemics: Hannah Arendt and her critics. *Dissent Magazine.* Retrieved from: https://www.dissentmagazine.org/democratiya_article/the-eichmann-polemics-hannah-arendt-and-her-critics. Accessed 11/18/19.

Gollwitzer, P. M. (1999). Implementation intentions: strong effects of simple plans. *American Psychologist*, *54*, 493.

Gollwitzer, P. M. (2012). Mindset theory of action phases. In P. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (pp. 526-545). London: Sage Publications.

Gollwitzer, P. M., Heckhausen, H., & Steller, B. (1990). Deliberative and implemental mind sets:Cognitive tuning toward congruous thoughts and information. *Journal of Personality and Social Psychology*, *59*, 1119.

Goodwyn, W. (2006). Former Enron Chairman Blames Others for Collapse. *NPR.* Retrieved from https://www.npr.org/templates/story/story.php?storyId=5361073.

Gray, K., & Schein, C. (2012). Two minds vs. two philosophies: Mind perception defines morality and dissolves the debate between deontology and utilitarianism. *Review of Philosophy and Psychology, 3*, 405-423.

Gray, K., & Wegner, D. M. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review, 14*, 7-16.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality.

Psychological Inquiry: An International Journal for the Advancement of Psychological Theory, 23, 101-124.

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.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105-2108.

Heckhausen, H. (1986). Why some time out might benefit achievement motivation research. In J. H. L. van den Bercken, T. C. M. Bergen, & E. E. J. De Bruyn (Eds.), *Achievement and task motivation* (pp. 7–39). Lisse, The Netherlands: Swets & Zeitlinger.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2009). Conservatives are more easily disgusted than liberals. *Cognition and emotion, 23*, 714-725.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33,* 315-365.

Knobe, J. & Burra, A. (2006). The folk concepts of intention and intentional action: A cross-cultural study. Journal of Cognition and Culture, *6*, 113-132.

Leslie, A., Knobe, J. & Cohen, A. (2006). Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science,* 17, 421-427.

Malle, B. F. (2010). Intentional action in folk psychology. In T. O'Connor and C. Sandis (Eds.), *A companion to the philosophy of action* (pp. 357-365). Chichester, UK: Wiley Blackwell.

Moore, J. & Slater, W. (2003). *Bush's Brain: How Karl Rove made George W. Bush presidential.* John Wiley & Sons, Inc.

Nadelhoffer, T. (2006). Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality. *Philosophical Explorations, 9*(2), 203-219.

Pallardy, R. (2018). Deepwater Horizon oil spill of 2010. *Encyclopedia Britannica.* Retrieved from https://www.britannica.com/event/Deepwater-Horizon-oil-spill-of-2010.

Pettit, D. & Knobe, J. (2009). The Pervasive Impact of Moral Judgment. *Mind & Language, 24*(5), 586-604.

Pramuk, J. (2018). Here's what the Cohen sentencing memos say about his efforts to help Trump's 2016 campaign. *CNBC.* Retrieved from https://www.cnbc.com/2018/12/07/cohen-sentencing-memo-details-efforts-to-help-trump-2016-campaign.html.

Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science, 29*, 1665-1678.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2012). P-curve: A key to the file-drawer. *Journal Experimental Psychology: General, 143*, 534–547.

Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science, 23*(1), 77-85.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science, 16*, 780-784.

Young, L., Cushman, F., Adolphs, R., Tranel, D., & Hauser, M. (2006). Does emotion mediate the relationship between an action's moral status and its intentional status? Neuropsychological evidence. *Journal of cognition and culture, 6*, 291-304.

Supplementary Materials

Experiment 1

Vignette

An oil rig explodes in the middle of the ocean, causing one of the most detrimental oil spills in history. The rig that exploded was owned and operated by drilling company **Atlantix** and installed by **Severton.** The rig had a weak concrete core and exploded once natural gas traveled through, killing 15 workers and injuring 26. Although owned by **Atlantix**, the rig was leased by oil company **PetroCorps**, who failed to initiate the rig's fail-safe designed to close the channel and prevent spills. As a result, millions of barrels of oil leaked into the sea and created a spill spanning four countries, causing billions of dollars in damage and contaminated waters, killing countless animals, and leading to massive unemployment. In this scenario, the question of culpability would quickly arise.

Atlantix owned the rig; **Severton** installed the malfunctioning rig; and **PetroCorps** made all decisions regarding installation.

Experiment 2a

Modified vignette

The Vice President of a company approached the Chairman of the Board and said, ‘We are thinking of starting a new program. It will help us increase profits, but it will also harm/help the environment.’ The Chairman of the Board **deliberated and decided**, ‘I don’t care at all about harming/helping the environment. I just want to make as much profit as I can. Let’s start the new

program.' The Vice President **planned and implemented** the new program. Sure enough, the environment was harmed/helped.

Experiment 2b

Modified vignette

As you requested, I looked into the new program you came up with and it will help us increase profits. It will also harm the environment'. The Chairman of the Board deliberated and then decided 'I don't care at all about harming/helping the environment. I just want to make as much profit as I can. Let's start the new program.' The Vice-President planned and implemented the new program. Sure enough, the environment was harmed/helped.

Original vignette (Knobe, 2003)

The Vice President of a company approached the chairman of the board and said, 'We are thinking of starting a new program. It will help us increase profits, but it will also harm the environment.' The chairman of the board answered, 'I don't care at all about harming/helping the environment. I just want to make as much profit as I can. Let's start the new program.' The Vice-President started the new program. Sure enough, the environment was harmed/helped.