It's not what you did, it's what you could have done[1]

Regan M. Bernhard[1,2], Hannah LeBaron[2], and Jonathan Phillips[2]
[1]Harvard University, [2]Dartmouth College

We are more likely to judge agents as morally culpable after we learn they acted freely rather than under duress or coercion. Interestingly, the reverse is also true: Individuals are more likely to be judged to have acted freely after we learn that they committed a moral violation. Researchers have argued that morality affects judgments of force by making the alternative actions the agent could have done instead appear comparatively normal, which then increases the perceived availability of relevant alternative actions. Across four studies, we test the novel predictions of this account. We find that the degree to which participants view possible alternative actions as normal strongly predicts their perceptions that an agent acted freely. This pattern holds both for perceptions of descriptive normality (whether the actions are unusual) and prescriptive normality (whether the actions are good) and persists even when what is actually done is held constant. We also find that manipulating the prudential value of alternative actions or the degree to which alternatives adhere to social norms, has a similar effect to manipulating whether the actions or their alternatives violate moral norms, and that both effects are explained by changes in the perceived normality of the alternatives. Finally, we even find that evaluations of both the prescriptive and descriptive normality of alternative actions explains force judgments in response to moral violations. Together, these results suggest that across contexts, participants' force judgments depend not on the morality of the actual action taken, but on the normality of possible alternatives. More broadly, our results build on prior work that suggests a unifying role of normality and counterfactuals across many areas of high-level human cognition.

[1] Correspondence for this article should be sent to Regan Bernhard, 13 Lowry Rd. Framingham, MA 01701, United States. Email: regan.bernhard@gmail.com. Funding: This work was supported by grant #61061 from The John Templeton Foundation. Draft version 2.0, 11/12/21. This paper has not been peer reviewed. Please do not copy or cite without author's permission. *Special thanks to Fiery Cushman for his ongoing help and support throughout the entire research process.*

1. Force and Norm

There is a clear relationship between normative judgments and assessments of force and freedom. For example, we are more likely to judge agents as morally culpable when we learn that they have acted freely rather than under duress or coercion (Darley & Shultz, 1990; Woolfolk, Doris, & Darley, 2006). Perhaps less intuitively, the reverse is also true: normative judgments influence evaluations of whether or not an agent acted freely (Harvey, Harris, & Barnes, 1975; Phillips & Knobe, 2009). Specifically, individuals are more likely to be presumed to have acted freely when engaging in moral violations than when committing morally neutral or morally good acts (Chakroff & Young, 2015; Phillips & Knobe, 2009; Young & Phillips, 2011).

To get a sense for this pattern, consider two cases from Phillips and Knobe (2009) in which a subordinate doctor at a hospital must follow the orders of the chief surgeon. In one case, the chief surgeon orders the subordinate doctor to prescribe medication, and the subordinate doctor, who dislikes the patient, reluctantly follows this order, realizing it will help the patient recover. Was the doctor forced to prescribe the medication? Participants typically judge that he was. But what if the chief surgeon ordered the doctor to prescribe medication that will instead harm the patient, and the doctor, who in this case likes the patient, again reluctantly does so? Was he forced in this second case? Most participants judge that he wasn't. A similar finding has been found across a range of cases, and the general pattern can be summarized in the following way: Given a fixed level of situational constraint, agents are more perceived as forced to take morally neutral or good actions but are not perceived as forced to commit moral transgressions (Phillips & Knobe, 2009; Mandelkern & Phillips, 2018; Young & Phillips, 2011). The critical question, which we address in this paper, is why such an effect occurs.

Importantly, judgments of force are not unique in being affected by the perceived moral status of an action. People's moral judgments similarly influence their intuitions about whether an action *caused* some further outcome, whether an agent *acted intentionally*, whether an agent *did something* or merely *allowed it* to happen, and many other judgments as well (Cushman, Knobe, & Sinnott-Armstrong, 2008; Hitchcock & Knobe, 2009; Knobe, 2003). Thus, to answer the question of why moral judgments influence judgments of force, it is important to keep in mind that the answer may not be specific to force judgments, but instead may arise from a more general effect of morality on a broad range of non-moral judgments (Phillips & Knobe, 2018, Phillips, Luguri, & Knobe, 2015; although see Hindriks, 2014, which argues against a unified treatment).

Two broad families of explanation have been offered. On the one hand, a number of researchers argue that morality influences these judgments via motivated moral reasoning. In this case, force judgments (as well as judgments of causation, intentionality, etc.) are altered, either consciously or unconsciously, such that they support an initial moral judgment of the action (e.g., Alicke, 2008; Nadelhoffer, 2004). On the other hand, researchers have argued that counterfactual thinking drives the effect of morality on these judgments. Specifically, this work suggests that we judge agents as acting freely, intentionally, or causally when committing moral transgressions because we are more likely to consider possible alternatives when evaluating morally bad *vs*. morally neutral or good actions (e.g., Halpern & Hitchcock, 2014; Knobe & Szabo, 2013; Phillips & Knobe, 2009; Phillips, Lugari, & Knobe, 2015; Young & Phillips, 2011).

To make progress on this question, the present research considers judgments of force as a case study that can help to shed light on the more general debate. Specifically, we argue that counterfactual relevance accounts make a number of novel predictions, which we test across four studies. The central idea we pursue is that according to such counterfactual theories, morality should not be special in its effect on judgments of force. If morality influences judgments of force by changing the perceived normality of alternative actions, then other factors that also affect the perceived normality of alternative actions should exhibit similar effects. Likewise, assessments of normality should predict changes in judgments of force in much the same way as assessments of the moral value do, since changes in moral value will result in changes in the normality of an action. Before turning to our experimental tests of these predictions, we briefly review the two different families of explanations, and illustrate why the counterfactual account—but not the motivated moral reasoning account—makes these predictions.

1.1. Motivated moral reasoning accounts

Motivated moral reasoning accounts argue that people's typical competency for making judgments of force, causation, intentionality, etc. can be distorted by their moral judgments. According to such accounts, when individuals respond to questions like "Was the agent forced?" or "Did the agent cause the outcome?" their answers do not reflect their underlying competencies, but instead reflect a motivation to have these judgments align with their moral assessment of the situation. While the specifics of these accounts vary, many rest on the claim that individuals seek to blame or punish agents for bad outcomes (Alicke, 2008; Alicke, Rose, & Bloom, 2011; Adams & Steadman, 2004; Clark et al., 2014; Clark, Baumeister, & Ditto, 2017; Clark, Winegard, & Shariff, 2019; Malle & Nelson, 2003; Mele, 2003; Nadelhoffer 2004; Nadelhoffer, 2006). Potentially motivated by an emotional response to harm (Nadelhoffer, 2004), individuals' desire to hold somebody responsible for this bad outcome shifts non-moral perceptions of the action (e.g. whether the agent acted freely, intentionally, or caused the outcome) to validate ascriptions of blame (Alicke, 2008) or moral responsibility (Clark et al., 2014). In support of this account, researchers have shown, for example, that when agents violate moral norms, their actions are judged to be more causal of bad outcomes than good outcomes (Alicke, Rose, & Bloom, 2011).

Belief in free will has been argued to be similarly influenced by a desire to punish, blame, or hold agents responsible for bad outcomes (Clark et al., 2014; Clark, Baumeister, & Ditto, 2017; see Clark, Winegard, & Shariff, 2019, for a meta-analysis). Reading about an agent's immoral actions not only increases study participants' perceptions that the agent acted freely, but also increases beliefs in metaphysical free will (Clark et al., 2014). Importantly, both of these effects are mediated by the degree to which participants would like to punish the transgressor. These effects are consistent both when participants are making judgments about hypothetical vignettes and ostensibly real events, and persist when free will beliefs are measured both directly and indirectly (Clark et al., 2014). Likewise, individuals who have a stronger tendency to moralize behaviors and to assign blame for moral transgressions, generally have stronger beliefs in free will (Everett et al., 2021). For example, political conservatives tend to report higher free will beliefs than liberals, and this difference is mediated by conservatives' stronger belief that individuals should be held morally responsible for their actions.

Critically, because such accounts focus on moral blame, punishment, or condemnation, the predictions of these accounts do not go beyond instances of *moral* transgressions. This is one of the key differences between motivated moral reasoning accounts and counterfactual relevance explanations described below.

1.2 Counterfactual relevance accounts

An alternative to the motivated moral reasoning perspective starts by pointing out that many of the non-moral judgments that are impacted by morality, share a particular feature: They all require reasoning about alternative possibilities (Phillips & Knobe, 2018; Phillips, Luguri, & Knobe, 2015). Judgments of force and freedom are particularly closely linked to considerations of counterfactual alternatives. In the most fundamental sense, one is forced to act specifically when there is no possible alternative action available. In fact, empirical research demonstrates that force judgments are heavily influenced by whether or not an agent was physically capable of taking another action (Woolfolk, Doris, & Darley, 2006). Likewise causal judgments, judgments of intentionality, and judgments of whether an agent did something or merely allowed it to happen, also all rely on consideration of counterfactual alternatives (e.g. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2020; Knobe, 2010; Phillips, et. al., 2015).

The idea that the effect of morality on judgments of force, causation, and intention is a *feature* of the decision-making process, stands in stark contrast to the motivated moral reasoning accounts, in which the effect of morality on all of these non-moral judgments is seen as an *error* or bias (for discussion, see, e.g., Halpern & Hitchcock, 2015; Kahneman & Miller, 1986; Knobe & Szabo, 2013; Kominsky & Phillips, 2019; Pettit & Knobe, 2009). Counterfactual thinking accounts argue that morality influences these non-moral judgments by affecting the perceived relevance of counterfactual alternatives. These morality-driven changes in the relevance of counterfactual alternatives subsequently affect the degree to which an agent's actions are seen as causal, intentional, forced, etc. Consistent with this view, some have argued that alternatives are considered more relevant if they replace an immoral action with something morally good or neutral (Halpern & Hitchcock, 2014; Knobe & Szabo, 2013). In fact, when asked, people do report considering counterfactuals in which morally good, rather than morally bad things occur (McCloy and Byrne, 2000; N'gbala & Branscombe, 1995).

One key piece of evidence for these counterfactual thinking accounts is that violations of any norm—whether it is descriptive (what is probable or usual) or prescriptive (what is valuable or good)—seem to have a similar effect on many of these non-moral judgments (Bear & Knobe, 2017; Kominsky, Phillips, Gerstenberg, Lagnado, & Knobe, 2015; Monroe & Ysidron, 2021).

For example, whether an agent is considered to have caused an outcome is influenced by both the morality of the action and how typical it is (Halpern & Hitchcock, 2015; Icard, Kominski, & Knobe, 2017; Kominski & Phillips, 2019; Kominski, et al., 2015; Phillips, et al., 2015). Likewise, violations of social/conventional norms (Uttich & Lombrozo, 2010) and moral norms (Knobe, 2003; Pettit & Knobe, 2009) have the same impact on whether or not an agent was viewed as having acted intentionally. Finally, agents are perceived as acting more freely when they engage in an action that differs from their typical behavior than when they engage in a typical action, even when both the normal and unusual actions result in the same bad outcome (Fillon, Lantian, Feldman, N'gbala, 2021). One unifying explanation for each of these findings is that the normality of the action taken affects the degree to which counterfactual actions are considered relevant and, consequently, perceptions of how causal or intentional the actual actions were (Phillips & Knobe, 2018; Phillips, Lugari, & Knobe, 2015). Importantly, such findings are left unexplained by motivated moral reasoning accounts which rest on the idea that these non-moral judgments are influenced by the view that norm violators are responsible, blameworthy, and punishable for moral transgressions in particular.

There is some evidence that the effect of morality on force judgements, in particular, is a consequence of the effect of morality on the relevance of counterfactual alternatives. When making force judgments about moral violations, possible alternatives are viewed as more relevant, available, or possible because these alternatives involve actions that were morally better than what was actually done (Phillips, Lugari, & Knobe, 2015). Further, this research has found that individuals are more likely to believe an agent could have done something else, and that considering these alternative possibilities is more relevant when the agent engaged in a moral transgression than when the act was morally good or neutral.

1.3. The present research

Unlike motivated moral reasoning accounts, counterfactual relevance accounts predict that moral norms should *not* be special in their effect on judgments of force and freedom, and that moral norms only change judgments of force by changing the perception of the alternative actions available to the agent who acted immorally. We test four key predictions that arise from this explanatory structure:

Hypothesis 1: According to counterfactual relevance accounts, violating a moral norm is merely one example of the ways in which an action can be counternormative (and thereby make alternative actions more relevant). If this is correct, then we should see comparable force judgments in response to scenarios involving *non-moral* norm violations. In Experiments 1, 2, and 3, we evaluate this novel prediction by asking if judgments of force are similarly affected by changes in whether an action or its alternatives conform to or violate prudential norms (whether the action would be rational to do) and social norms (whether the action is typically done by others).

Hypothesis 2: Counterfactual relevance accounts argue that changes in the moral status of an action affect judgments of force *not* because they change our perception of the actual action that was done, but because they change our perception of the alternative actions available to the agent. To date, this critical aspect of counterfactual accounts has not been tested. We evaluate this claim by conducting experiments in which the actual action the agent does is held fixed, but we directly manipulate the perceived relative normality of the available alternative actions (Experiments 1 and 3).

Hypothesis 3: Counterfactual relevance accounts posit that the relevance of alternative actions depends not just on the degree to which these alternatives adhere to prescriptive norms (how good they are), but how well they adhere to descriptive norms (how typical they are) as well. We test this prediction in Experiments 1-4 by asking if force judgments are predicted by both evaluations of the value of alternative actions and evaluations of the alternatives' unusualness.

Hypothesis 4: Counterfactual relevance accounts propose that the effect of morality on non-moral judgments should be explainable in the same way as other kinds of norm violations. Therefore, even for moral norm violations specifically, the "unusualness" of potential alternative actions should mediate the effect of the moral valence of the actual action on force judgments. We evaluate this fourth novel prediction in Experiment 4.

Across all four experiments, we find robust evidence in favor of counterfactual relevance accounts of the impact of normality on judgments of force.

2. Experiment 1: The effect of prudential value on force judgments

As described above, counterfactual relevance accounts suggest that agents are perceived to have acted freely when possible alternative actions are seen as more normal than the actual action, and consequently relevant to consider. If the relative normality of alternatives drives force judgments, rather than the morality of the *actual* action, we would expect that agents will be perceived as acting freely whenever alternative actions are introduced that are at least as normal as the actual action. Importantly, the relative normality of alternative actions should affect force judgments outside of the moral domain in the same way they do within it (Hypothesis 1). In Experiment 1, we test this by asking participants to make force judgments in response to cases where possible alternative actions do or do not violate prudential (rather than moral) norms. We hypothesize that participants are more likely to view agents as acting freely when there is at least one possible alternative action of equivalent prudential value to the actual action. On the other hand, when all of the available alternatives require violating prudential norms, we predict that participants will increasingly view agents as forced to take the actions they did.

Counterfactual relevance accounts also predict that both the prescriptive and descriptive normality of possible alternative actions will affect force judgments (Hypothesis 3). To investigate this prediction, participants made judgments of the normality of actions the agent could have taken but didn't. Importantly, we ask participants to make judgments about both how good the alternatives are (to measure the prescriptive normality of the alternatives) and how unusual the alternatives are (measuring the descriptive normality of the alternatives). We predict that both the value and likelihood of alternatives will independently predict force judgments.

2.1. Methods

To test the effects of prudential norm violations on force judgments we conducted two experiments. The basic study design was similar for both experiments. All participants were presented with the following scenario:

> *Imagine there is a reality TV show where contestants will spend one month alone on a desert island and the challenge is to survive. Contestants are not allowed to bring anything but the clothes they are wearing. However, the producers will begin the show by offering each contestant a "survival pack" with three items. Each contestant is allowed to choose exactly two of the items to bring on the island. The contestant must reject one of the items and leave it behind. The contestants' decisions about what to keep and what to reject is very important. In the past, people have not made it to the end of the month if they haven't kept the right things. Imagine that you are watching one contestant, Joe, open the survival pack on TV. You will find out what is in the survival pack and what Joe decides to reject. You will then be asked a few questions about Joe's decision.*

In both experiments, participants were in one of two conditions. In the "normal alternative" condition participants learned that the survival pack contained two items that would not be useful to have on a desert island (e.g. a ring and a stuffed animal) and one item that would be useful to have on a desert island (e.g. an axe). Participants were then told that the contestant decided to reject one of the less useful items (e.g. the ring) from the survival pack and leave it behind. Critically, in this condition there was an alternative action the contestant could have taken (rejecting the other less useful item, e.g. rejecting the stuffed animal instead) that would have been of equivalent prudential value. In the "abnormal alternative" condition, participants were told that the survival pack had only one item that would not be useful to have on a desert island (e.g. a ring) and two items that would be very useful to have on a desert island (e.g. an axe and a fishing pole). Once again, participants learned that the contestant rejected the less useful item (the ring). However, this time, there is no other option available of equivalent prudential value. The contestants' other two choices (e.g. rejecting the axe or the fishing pole) would require leaving something useful behind.

After participants read the scenario and the contestants' decision, they were asked to make a force judgment. To do this, in Experiment 1a, participants then rated their agreement to statements such as "Seems like [the contestant] had to reject the [rejected item]" on a 1-5 scale (where 1 is Strongly Disagree and 5 is Strongly Agree). In this within-subjects version of the experiment, each participant responded to eight trials, four in each condition.[2]

[2] Data for Experiment 1a was originally collected as a pilot study for another project. Therefore, there are additional details in the experimental design that are irrelevant for the present project. These additional design details are described in the Supplementary Information.

In Experiment 1b, participants were instead asked to respond to the question, "How strongly do you agree that [the contestant] had to reject the [item that was rejected]?" Subjects indicated their response using a slider on a scale ranging from 0 (Strongly Disagree) to 100 (Strongly Agree). In this between-subjects experiment, participants responded only to one trial, and were placed in either the normal alternative or abnormal alternative condition. In Experiment 1b we also measured participants' perceptions of the normality of possible alternative actions. To measure participants' judgments about the degree to which the available alternative actions adhere to prescriptive norms, participants were asked, "How strongly do you agree that it would have been a good idea for [the contestant] to reject the [one of the items that was *not* rejected] instead?" Participants responded to this question once for each of the unrejected items from the survival pack. To measure participants' judgments about the degree to which the available alternative actions adhere to descriptive norms, participants were asked "How strongly do you agree that it would have been unusual for [the contestant] to reject the [one of the items that was *not* rejected]?" Again, participants responded to this question once for each of the unrejected items. Participants responded to each question using a slider that ranged from 0 (Strongly Disagree) to 100 (Strongly Agree). These questions were presented in random order across participants. Importantly, in both experiments, in the normal alternative and abnormal alternative conditions, the actual action (rejecting a specific useless item) is consistent. Thus, any differences in force judgment between conditions must depend on features of the counterfactual actions rather than of the actual action.

In Experiment 1a we collected data from 26 subjects (34% female) and in Experiment 1b, we collected data from 102 subjects (42% female). All data collection was completed through Amazon Mechanical Turk. Recruitment was automated with TurkPrime (www.turkprime.com) to prevent repeat participation across each of the studies presented here. Recruitment was limited to participants living in the United States, who had participated in at least 1,000 previous studies, and who had been approved on 95% of those studies. Experiment 1a took approximately 20 minutes to complete and participants were paid $2.00 for their participation. Experiment 1b took approximately 2 minutes and participants were paid $0.25.

In Experiment 1b, participants also completed three attention check questions (see SI for details) to ensure high quality data. Participants who did not respond correctly to all three questions were eliminated from further analyses (21 participants). In addition, we excluded participants who did not complete the entire study (6 participants). After applying these exclusion criteria, data from 75 participants were analyzed for Experiment 1b. These and all other experimental methods described in this paper were approved by the Harvard University Committee on the Use of Human Subjects.

2.2. Results

In Experiment 1a we tested the effect of condition (normal vs. abnormal alternative) on force judgments using a linear mixed-effects regression implemented with the *lme4* package in R

(Bates, Maechler, Bolker, & Walker, 2015). We began by predicting force judgments with a model that included condition and a random intercept for subject. We then compared this model to a reduced model which excluded condition. Model comparisons were conducted using the ANOVA function in R. We found that our model that included condition predicted force judgments significantly better than the reduced model (X^2 (1, *N* = 26) = 46.80, *p* < .0001, $b_{condition}$ = 1.09; Figure 1A). In other words, participants were much more likely agree that contestants had to take the action they did when the other two items in the survival pack were of high prudential value (abnormal alternative condition; *M* = 3.85, *SD* = 1.36) than when the survival pack contained one other low-value item (normal alternative condition; *M* = 2.56, *SD* = 1.13).

Similarly, in Experiment 1b a linear model revealed that condition significantly predicts subjects’ force judgments (*b* = 23.98, *t*(73) = 4.35, *p* < .0001; Figure 1A). Once again, force judgments were higher in the abnormal alternative condition (*M* = 95.83, *SD* = 8.49, *N* = 35) than in the normal alternative condition (*M* = 71.85, *SD* = 31.63, *N* = 40). Thus, across both experiments we found that participants were less likely to perceive agents as forced when there were alternative actions of equivalent prudential value the agents could have taken instead.

Individual participants’ judgments of the degree to which alternative actions violated prescriptive and descriptive norms also predicted force judgments. Recall that in Experiment 1b, participants were asked to rate how strongly they agreed that each alternative action would have been unusual (unusualness judgement). Between each participant’s two unusualness judgements, we identified the alternative action that was rated as being *least* unusual. Because we hypothesized that subjects are more likely to perceive an agent as forced when the alternative actions violate descriptive norms (are more unusual), we anticipate that the unusualness judgments of the *least* unusual alternative action would predict force judgments (the more unusual the least unusual action is, the more likely agents are perceived as forced). Likewise, participants also rated how strongly they agreed that each alternative action would have been a good idea to do instead (value judgments). Here we identified the alternative action that was rated as being the best idea. Because we hypothesized that subjects are more likely to perceive an agent as forced when the alternative actions violate prescriptive norms (are not a good idea), we anticipate that participants’ value judgments of the *best* alternative action would *negatively* predict force judgments (the better the best alternative is, the less likely agents are perceived as forced). We refer to the best alternative and the most unusual alternative as the “critical alternatives”.

To test these predictions, subjects’ lowest unusualness judgment and highest value judgment were entered into a single linear model predicting force judgments. As expected, the lowest unusualness judgments (unusualness judgments of the critical alternative) significantly predicted force judgments, when controlling for value judgments (*b* = 0.22, *t*(72) = 2.99, *p* < .005; Figure 1B). Similarly, the highest value judgments (value judgments of the critical alternative) significantly negatively predicted force judgments when controlling for unusualness judgments (*b* = -0.49, *t*(72) = -5.82, *p* < .0001; Figure 1C). In other words, the more participants perceived the alternative actions as being unusual and the less they perceived alternatives as being a good idea, the more likely they were to view the contestants as forced to take the actions they did. Interestingly, when we include condition in the model we see that controlling for unusualness and value judgments eliminates the effect of condition on force judgments (b = -4.40, $t(71) = -0.79$, $p = .43$). This suggests that the effect of condition on force judgments may be largely explained by subjects’ perceptions of the value and unusualness of the alternatives. The effect of value and unusualness judgments persists, even when controlling for condition (unusualness: $b = 0.25$, $t(71) = 2.96$, $p < .005$; value: $b = -0.51$, $t(71) = -5.84$, $p < .001$).

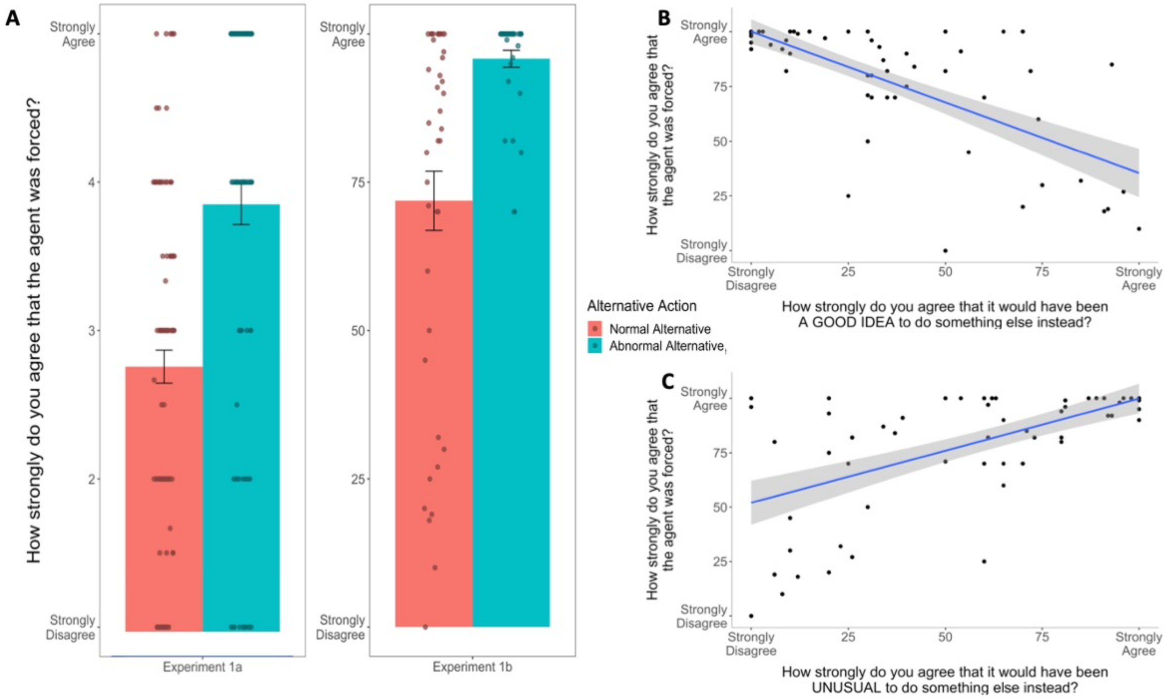

Figure 1*: *Experiment 1a and 1b Results. A: Participants’ average force judgments when there was a normal alternative action available vs when there were only abnormal alternative actions available. Force judgments were reported on a 1-5 scale in Experiment 1a, and on a 0-100 scale in Experiment 1b. B: Participants’ force judgments predicted by the degree to which they viewed the critical alternative action as a good idea to do instead (value judgments). C: Participants force judgments predicted by the degree to which they viewed the critical alternative action as unusual (unusualness judgments). Error bars on bar charts represent standard errors. Shaded areas in scatter plots represent 95% CI.

2.3 Discussion

In Experiment 1, we found that the degree to which participants viewed the agent as forced was driven by the availability of alternative actions of equivalent prudential value, and participants’ beliefs that these alternatives adhered to descriptive and prescriptive norms. These findings provide support for the critical role of counterfactual thinking in force judgments in several important ways. First, in support of Hypothesis 1, we found that the normality of alternative actions affects force judgment outside of the moral domain. In addition, in this study, the agent takes the *same* action (rejecting a low value item) in both conditions. Therefore, in support of Hypothesis 2, we found that rather than being driven by an evaluation of the actual action, force judgments in this experiment must be based on what alternatives are available. Finally, in support of Hypothesis 3, we found that evaluations of the prescriptive normality of possible alternative actions doesn't solely drive force judgments. Rather, how typical, or the degree to which alternatives adhere to descriptive norms, also predicts force judgments.

An important distinction between this study and force judgments in response to moral dilemmas is that in cases like that of the doctor described in the Introduction, participants are making force judgments in response to norm violations, or morally bad actions. Conversely, in our Experiment 1, the contestant's actions were always good choices - those of high prudential value. One could argue then that when making force judgments in response to neutral or good actions, we appeal to the availability of normative counterfactual actions, but when making force judgments in response to moral transgressions or other types of norm violations, our judgments are based on an evaluation of the actions themselves. Two address this claim, in Experiment 2, participants make force judgments in response to scenarios in which the actual action involves norm violations. In addition, to further test the relationship between the normality of counterfactual actions and force judgments outside the moral domain, in Experiment 2, subjects make force judgments in response to scenarios in which agents violate social norms.

3. Experiment 2: Force judgments in response to social norm violations

In Experiment 2 we asked participants to make force judgments in response to an agent engaging in an action that either adheres to or violates social norms. As in Experiment 1, we also asked participants to assess how unusual and good alternative actions are. We expected that, once again, participants' force judgments will be driven by the degree to which the alternative actions are viewed as normal. Importantly, we expect this to be true whether the agent's actual action adheres to or violates social norms.

3.1. Methods

In Experiment 2, data was collected from a total of 908 participants (47% female). Participants were recruited from Amazon Mechanical Turk and were each paid $0.25 in compensation. Recruitment was limited to participants living in the United States, who had participated in at least 1,000 previous studies, who had been approved on 95% of those studies, and who had not participated in related studies in the past. We excluded participants who did not complete the entire study (64 participants). After applying these exclusion criteria, data from 844 participants were analyzed. The sample sizes, exclusion criteria, experimental paradigm, and analysis plan were all preregistered (AsPredicted #37496).

Prior to beginning this preregistered experiment, we conducted several pilot studies to determine whether we could design vignettes in which we saw an effect of social norm violations on force judgments. Once we determined that we were able to do so, we shifted our focus to our central motivation, which is evaluating what factors drive this effect. To this end, in Experiment 2, participants responded to one of three scenarios describing an agent who engaged in either a norm congruent (described as the action that most of the other people in the scenario take) or norm violating action. For the scenarios in which the agent’s actual action adhered to a social norm, the possible alternative action required violating that social norm. Alternatively, for scenarios in which the agent’s actual action violated a social norm, the possible alternative action adhered to that norm. Thus, participants were placed in one of two conditions: the normal alternative condition (where the actual action violated a social norm, but the alternative adhered to that norm) and the abnormal alternative condition (where the actual action adhered to a social norm, but the alternative was a norm violation). For example, in the “lab” vignette, participants read the following:

Melissa is a scientist working in a lab with many different kinds of chemicals. The lab director requires that everyone in the lab wear gloves to protect themselves from the chemicals. All of the other scientists wear latex gloves but there are also vinyl gloves available. One day, Melissa decides to wear the [latex/vinyl] gloves.

We measured participants’ force judgments by asking them to use a slider to indicate how strongly they agreed that the agent had to take the action they did on a scale ranging from 0 (Strongly Disagree) to 100 (Strongly Agree). Subsequently, we measured the degree to which participants viewed the non-chosen alternative action as adhering to both prescriptive and descriptive norms by asking them to rate the degree to which they perceived the alternative actions as a “good idea” and as “unusual”. Once again, these two follow-up questions were displayed in random order across participants.

3.2. Results

In this and all of the following Experiments, analyses occurred in two stages. In the first stage we tested the effect of condition (here, normal alternative vs. abnormal alternative) on force judgments. This analysis was conducted using a linear mixed-effects regression implemented with the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015). We began by predicting force judgments with a model that included condition and a random intercept for scenario. We then compared this model to a reduced model which excluded condition. Model comparisons were conducted using an ANOVA. We found that our model with condition predicted force judgments significantly better than the reduced model. Specifically, participants were more likely to agree that agents were forced when they adhered to social norms (but could have performed a social norm violating action instead; abnormal alternative condition, averaging across scenarios, (*M* = 45.54, *SD* = 31.50), than when they violated social norms (but could have performed a social norm adhering action instead; normal alternative condition, averaging across scenarios, *M* = 33.58, *SD* = 31.74).

The second stage of our analyses investigates the effect of participants’ perceptions of the normality of possible alternatives on force judgments. This analysis was similarly conducted using a linear mixed-effects regression implemented with the *lme4* package in R (Bates, Maechler, Bolker, & Walker, 2015). We began by predicting force judgments with a model that included value and unusualness judgments, and a random intercept for scenario. We then separately dropped value and unusualness judgments from the model comparing these reduced models to our original model. Again model comparisons were conducted using an ANOVA. We found that the original model performed significantly better than a model excluding unusualness judgments (X^2 (1, N = 844) = 88.98, $p < .0001$, $b_{unusualness} = 0.32$; Figure 2b). In other words, the more unusual participants viewed the alternative actions the more they perceived agents as forced to take the actions they did. We also found that the original model performed significantly better than the model excluding value judgments (X^2 (1, N = 844) = 13.86, $p < .0005$, b_{value} = 0.18; Figure 2c). Unexpectedly, the relationship between value judgments and force judgments was positive. In other words, the more strongly participants viewed the alternatives as a good idea, the more strongly they perceived agents as forced to take the actions they did. Finally, we found that adding condition to our original model did not significantly improve model fit (X^2 (1, N = 844) = 2.99, $p < .09$).

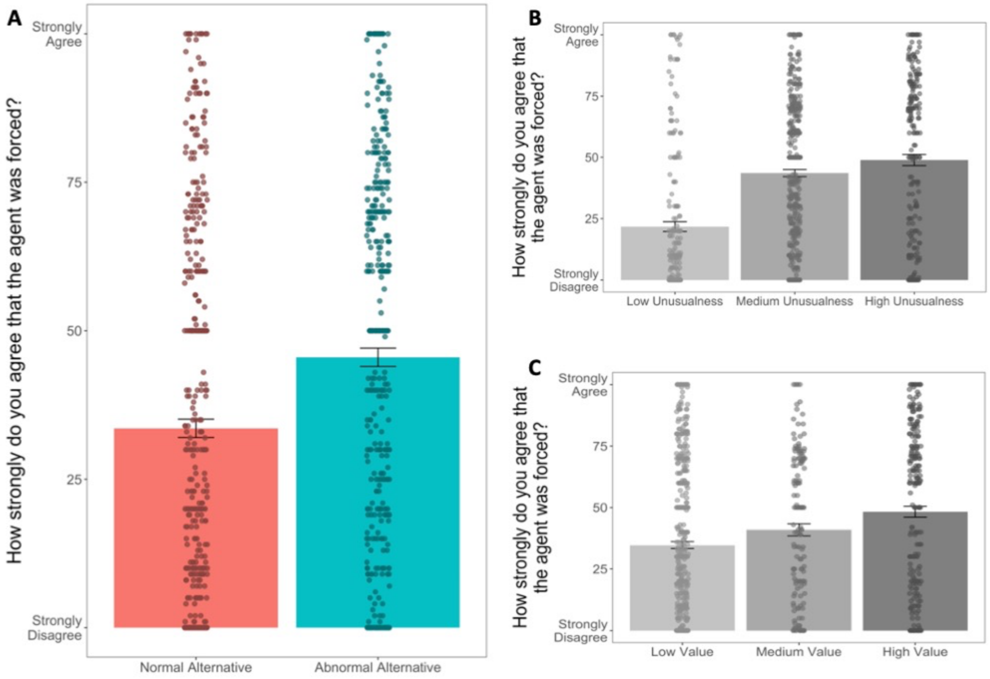

Figure 2*: *Experiment 2 Results. A: Participants’ average force judgments when there was a normal alternative action available and when there was only an abnormal alternative action available. B/C: Participants’ force judgments predicted by the degree to which they viewed the critical alternative action as unusual (Figure 2b) or a good idea to do instead (Figure 2c). For visualization purposes participants were grouped into three categories: Low Unusualness/Value are those participants whose unusualness/value ratings of the alternative fell in the bottom 25%, Medium Unusualness/Value are those whose unusualness/value ratings fell in the middle 50%, and High Unusualness/Value are those whose unusualness/value ratings fell in the top 25%. Error bars represent standard errors.

3.3 Discussion

In Experiment 2, we found that participants were more likely to view an agent as forced when the alternative action required violating a social norm than when the possible alternative action adhered to social norms. In support of Hypothesis 1, and consistent with our findings from Experiment 1, we found that whether actions and their alternatives adhered to norms affected force judgments outside of the moral domain - in this case in response to social norm violations. More centrally, and in support of Hypothesis 3, we found a clear relationship between participants’ evaluations of the unusualness of alternative actions and force judgments.

Unlike in Experiment 1, in this study, the normality of the alternatives across conditions was confounded with the normality of the actual actions. Recall that in the normal alternative condition, the actual action involved a norm violation, but in the abnormal alternative condition, the actual action was norm-congruent. Therefore, we are unable to ensure that our effects were truly a consequence of the normality of the alternative action or due to evaluations of the actual action. In Experiment 3, we address this concern by disentangling the effects of the normality of the actual action from the normality of the alternatives on force judgments.

4. Experiment 3: Force judgments in response to potential social norm violations holding the actual action fixed

Although the results of Experiment 2 suggest that subjects’ force judgments are driven by evaluations of possible alternative actions, that experiment does not control for the effect of the normality of the actual action. Therefore, in Experiment 3, we address this concern by holding the actual action fixed across conditions. The ambiguous relationship between value and force judgments identified in Experiment 2, also suggests that agents can be perceived as forced when the alternative actions violate descriptive norms alone, even when the alternative actions are perceived as good. That is, one might intuitively think that we view agents as forced when their alternative actions are in some way bad (either morally or socially). However, the Experiment 2 results suggest that alternative actions need not be bad to shift force judgments, they can simply be unusual. In Experiment 3 we test this claim by providing sets of possible actions in which the degree to which they adhere to social norms (the kinds of things people usually do) is disentangled from their value.

4.1 Methods

Data was collected through Amazon Mechanical Turk from 302 subjects for an original study (Experiment 3a) and from 905 subjects in a pre-registered direct replication (Experiment 3b; AsPredicted #46986). Recruitment was limited to participants living in the United States, who had participated in at least 1,000 previous studies, and who had been approved on 95% of those studies, and who had not participated in related studies in the past. Participants were paid

$0.25 for completing the 2 minute survey. In order to ensure quality data, participants completed three attention check questions after completing the main component of the experiment (see SI for details). Participants who did not correctly answer all three questions, or did not complete the entire study were eliminated from further analyses (56 participants excluded from Experiment 3a, and 140 participants excluded from Experiment 3b). Our final sample sizes were 246 subjects for Experiment 3a, and 765 participants for Experiment 3b. Prior to conducting Experiments 3a and 3b we ran several pilot studies to ensure that we were able to create scenarios in which possible alternative actions varied in their degree of unusualness, without changing how good they were perceived to be.

In this experiment participants read one of three scenarios (featuring an agent lifting weights, trying to cool down on a hot day, or taking a walk for more exercise). In each of these scenarios, an agent is required to perform an action and is given three different options for how to do so. For example:

Jacob is just starting out as a competitive bodybuilder. His coach tells him he needs to work on his upper body strength and that in order to maximize his strength gains he needs to spend several weeks lifting 50lb weights. The only things in Jacob's house that weigh 50lbs are: 50lbs on a weightlifting machine that gets jammed sometimes, 50lbs in rusty dumbbells, and a 50lb bag of dog food. Jacob decides to strengthen his arms by lifting 50lbs on a weightlifting machine that gets jammed sometimes.

In each scenario, some of the possible actions are socially normative (e.g. lifting weights on a weight lifting machine or lifting dumbbells for exercise) and some are unusual (e.g. lifting a bag of dog food for exercise). However, to try to mitigate any possible effects of value on participants' judgments, we described the normative actions as having some kind of flaw (e.g. the weight machine gets jammed sometimes or the dumbbells are rusty). Subjects responded to a single scenario that fell in one of two conditions. In the "normal alternative" condition, the agent's options included two descriptively normal actions and one unusual action (as in the example above). In the "abnormal alternative" condition, the agent had one descriptively normal option and two unusual options (e.g. lifting a bag of dog food or lifting a jug of water). Critically, the agent always chooses to take a descriptively normal action (e.g. lifting weights). Therefore, in the normal alternative condition, there is something else the agent could have done instead that adheres to social norms. Conversely, in the abnormal alternative condition, the agent's possible alternative actions are all unusual. Because the actual action is held fixed across both conditions, any relationship we see between possible actions and force judgment must be a product of the alternative actions taken, not of the actual action.

As in the previous experiments, we measured participants' force judgments by asking them to use a slider to indicate on a scale ranging from 0 (Strongly Disagree) to 100 (Strongly Agree) how strongly they agreed that the agent had to take the action they did. Once again, we subsequently measured the degree to which participants viewed the possible alternative actions as adhering to both prescriptive and descriptive norms by asking them to rate the degree to which they perceived the alternative actions as a “good idea” and as “unusual”. The order of these two questions was randomized.

4.2. Results

Once again, these analyses were done in two stages. In the first stage we used a similar model comparison approach as described in Experiment 2 to test the effect of condition on force judgments. Unlike in our previous experiment, however, here our model that included condition and a random intercept for scenario did not perform significantly better than a reduced model excluding condition (Experiment 3a: X^2 (1, N = 246) = 0.1, p = 0.75, $b_{condition}$ = 1.17; Experiment 3b: X^2 (1, N = 765) = 2.13, p = 0.14, $b_{condition}$ = 3.11; Figure 3a). Specifically, participants were no more likely to agree that agents were forced when there was no normative alternative available (Experiment 3a: M = 32.25, SD = 29.51; Experiment 3b: M = 30.65, SD = 29.46), then when there was a normative alternative available (Experiment 3a: M = 31.17, SD = 29.69; Experiment 3b: M = 34.22, SD = 30.41).

In the second stage of our analyses we used the same model reduction approach, as described in Experiment 2, to test the effects of unusualness and value judgments on force judgments. We began by identifying the alternative actions that each subject rated as being the least unusual and best idea (which we call the “critical alternatives”). We then used these ratings in a linear mixed effect regression to predict force judgments, with a random intercept for scenario. Finally, we separately dropped unusualness and value judgments from the model and used an ANOVA to compare these reduced models to the original model. In both Experiments 3a and 3b, we found that the original model performed significantly better than a reduced model excluding unusualness judgments (Experiment 3a: X^2 (1, N = 246) = 13.38, p < .0005, $b_{unusualness}$ = 0.27; Experiment 3b: X^2 (1, N = 765) = 43.78, p < .0001, $b_{unusualness}$ = 0.30; Figure 3b). However, we found that the original model did not perform significantly better than a reduced model excluding value judgments (Experiment 3a: X^2 (1, N = 246) = 0.48, p = 0.49, b_{value} = 0.06; Experiment 3b: X^2 (1, N = 765) = 2.50, p = .11, b_{value} = 0.08; Figure 3c). In Experiment 3b, but not in Experiment 3a, we found that adding condition into our original model did significantly improve model fit (Experiment 3a: X^2 (1, N = 246) = 0.35, p = 0.55; Experiment 3b: X^2 (1, N = 765) = 17.10, p < .0001).

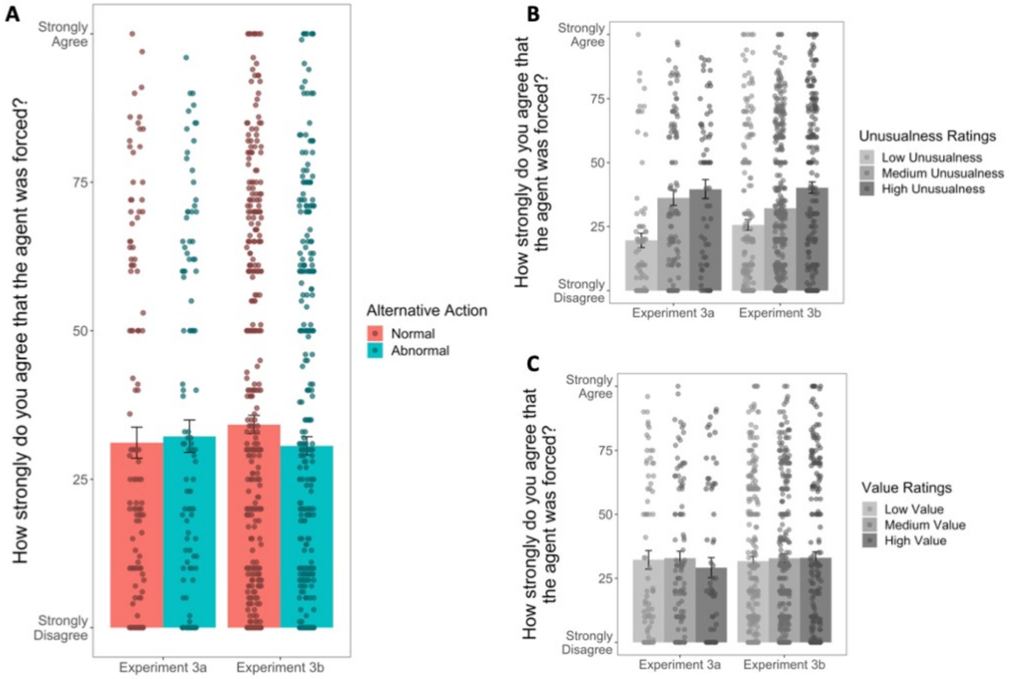

Figure 3*: *Experiment 3a and 3b Results. A: Participants' average force judgments when there was a normal alternative action available and when there was only abnormal alternative actions available. B/C: Participants' force judgments predicted by the degree to which they viewed the critical alternative action as unusual (Figure 3b) or a good idea (Figure 3c) to do instead. For visualization purposes participants were grouped into three categories: Low Unusualness/Value are those participants whose unusualness/value ratings of the critical alternative fell in the bottom 25%, Medium Unusualness/Value are those whose unusualness/value ratings of the critical alternative fell in the middle 50%, and High Unusualness/Value are those whose unusualness/value ratings of the critical alternative fell in the top 25%. Error bars represent standard errors.

4.3 Discussion

While we found no effect of condition on force judgments, we continue to find that the degree to which alternative actions are seen as adhering to descriptive norms predicts perceptions that an agent was forced (Hypothesis 3). Critically, in each scenario in this study, regardless of the available alternatives, the agent chose the same actual action. Therefore, consistent with Hypothesis 2, force judgments must depend on evaluations of the alternative. Interestingly, as we found in Experiment 2, it is not necessary for participants to perceive an agent as having a high value (good) alternative in order to see the agent as having acted freely. Rather, consistent with Hypothesis 3, it is sufficient that the agent has at least one possible alternative action that is perceived as common or normal.

In conjunction, the results from Experiments 1-3 suggest that force judgments of actions outside of the moral domain depend not on evaluations of the action taken, but on evaluations of the normality of possible alternative actions. And, moreover, that participants evaluate both the prescriptive *and* descriptive normality of possible alternatives when making force judgments. However, one could still argue that moral transgressions are a special class of actions, in which evaluations of the transgression supersede other considerations when making force judgments. To test this claim, in Experiment 4, we evaluate the relationship between perceptions of the normality of possible alternatives and force judgments in scenarios in which agents engage in morally wrong versus morally neutral acts.

5. Experiment 4: Force judgments in response to moral norm violations

In Experiment 4, we apply our general design to the original demonstration that morality affects judgments of force (Phillips & Knobe, 2009). In the original study, the authors found that participants were more likely to view the agent as forced when they engaged in a morally neutral action than when they committed a moral transgression. Moreover, they found that when the agents committed a moral transgression, participants more strongly believed they had the option of not doing so, then when their actual action was morally neutral. These findings provide some initial evidence that participants' force judgments in response to moral violations depend on perceptions of available alternatives. Here we build on these results by testing Hypothesis 4, that the normality of alternative actions is a source of this asymmetry even in the moral domain. In other words, the morality of an action affects force judgments by shifting the perceived normality of possible alternatives. Importantly, the counterfactual relevance accounts of force judgment presented here argue that both the prescriptive and descriptive normality of possible alternatives should affect force judgments, even in the moral domain.

5.1. Methods

Data was collected through Amazon Mechanical Turk from 214 subjects (39% female). As with all previous studies, recruitment was limited to participants living in the United States, who had participated in at least 1,000 previous studies, and who had been approved on 95% of those studies, and who had not participated in related studies in the past. Participants were paid $0.25 for completing the 2-minute survey. Participants who did not complete the entire study were eliminated from further analyses (11 participants), leaving a final sample size of 203 participants.

In this study participants responded to the same vignettes used in Experiments 1 and 2 of Phillips and Knobe, 2009. Like this prior work, participants were presented with either a morally good/neutral or morally bad version of one of two scenarios. For example, in this morally bad version of the ship captain scenario, participants read the following:

While sailing on the sea, a large storm came upon a captain and his ship. As the waves began to grow larger, the captain realized that his small vessel was too heavy, and the ship would flood if he didn't make it lighter. The only way that the captain could keep the ship from capsizing was to throw his wife overboard. Thinking quickly, the captain took his wife and tossed her into the sea. While the captain's wife sank to the bottom of the sea, the captain was able to survive the storm and returned home safely.

In the morally neutral version of this scenario, the captain opts to throw his wife's expensive cargo overboard instead of his wife. As in the previous experiments, we measured participants' force judgments by asking them to use a slider to indicate on a scale ranging from 0 (Strongly Disagree) to 100 (Strongly Agree) how strongly they agreed that the agent had to take the action they did. Unlike in our Experiments 1-3, in this experiment no specific alternative actions were provided. Therefore, to measure the degree to which participants viewed possible alternative actions as adhering to prescriptive norms we asked them to indicate how strongly they agreed that it "would have been a good idea to do something else instead." To measure the degree to which participants viewed possible alternative actions as adhering to descriptive norms, we asked them to indicate how strongly they agreed that it "would have been unusual to do something else instead."

5.1. Results

Following the two-stage analytic procedure described in Experiments 2 and 3 we began by testing the effect of condition on force judgments. Replicating the results from Phillips & Knobe (2009), we found that subjects were more likely to view an agent as forced when engaging in a morally neutral or good act (M = 85.0, SD = 23.32) than when committing a moral violation (M = 36.02, SD = 36.89). A model including condition and a random intercept for scenario predicted force judgments significantly better than a reduced model excluding condition (X^2 (1, N = 203) = 103.4, p < .00001, $b_{condition}$ = 49.05; Figure 4a).

In our second stage of analyses, we tested the effect of unusualness and value judgments on force judgments. We began by predicting force judgments with a model that included value and unusualness judgments, and a random intercept for scenario. We then separately dropped value and unusualness judgments from the model, comparing these reduced models to our original model. Consistent with the findings presented in Experiments 1-3, participants' evaluations of the normality of the alternative actions significantly predicted force judgments. We found that our original model performed significantly better than reduced models that excluded either value judgments (X^2 (1, N = 203) = 10.03, p < .005; Figure 4b) or unusualness judgements (X^2 (1, N = 203) = 75.88, p < .00001; Figure 4c). In other words, the more participants agreed that possible alternative actions would have been a good idea, the less they viewed the agents as forced to take the actions they did (b_{value} = -0.19). Conversely, the more participants agreed that possible alternative actions would have been unusual, the more they viewed the agents as forced to take the actions they did ($b_{unusualness}$ = 0.62).

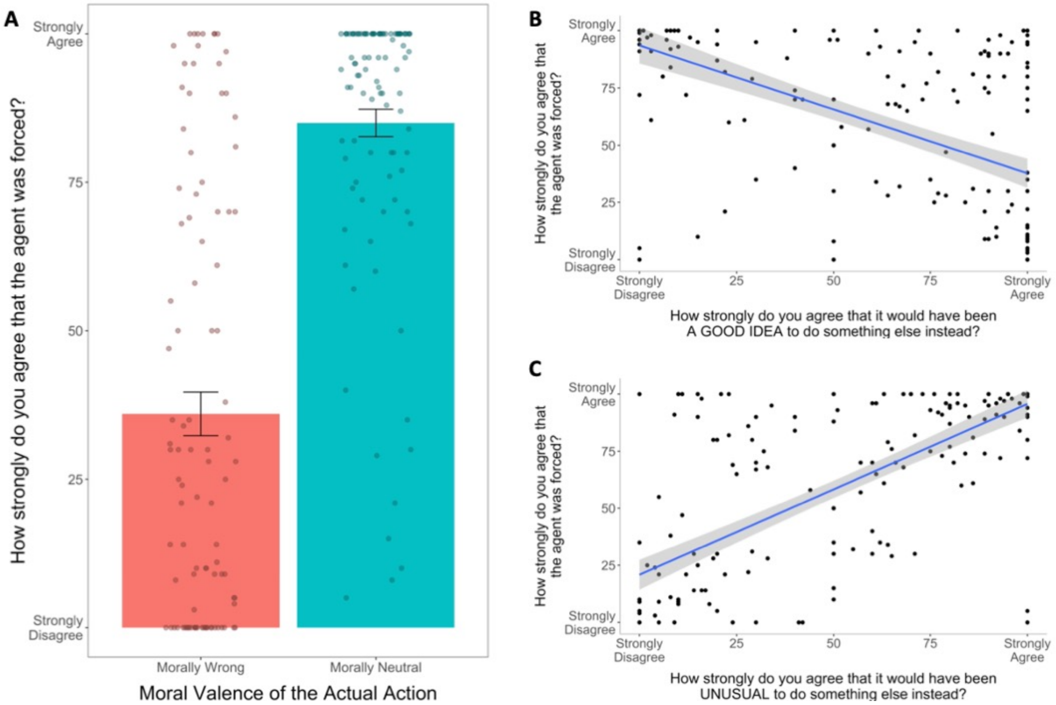

Figure 4*: *Experiment 4 Results. A: Participants' average force judgments when the agent engages in a morally neutral or morally wrong action. B/C: Participants' force judgments predicted by the degree to which they viewed the critical alternative action as unusual (Figure 4b) or a good idea (Figure 4c) to do instead. Error bars on bar charts represent standard errors. Shaded areas in scatter plots represent 95% CI.

That unusualness judgments so strongly predicted force judgments was especially surprising given that participants are making force judgments in a moral domain, where the value of an action and its alternatives is clearly relevant, but the likelihood seems less so. To probe these findings further, we performed two exploratory analyses. First, to compare the relative strengths of unusualness and value judgments as predictors of force judgment we conducted a Wald test using the *linearHypothesis* function in R (Fox & Weisberg, 2018). Perhaps counterintuitively, we found that unusualness judgments were a significantly better predictor of force judgments than value judgments (F(1, 199) = 13.44, $p < .001$). Second, our results thus far suggest that condition (whether or not the agent engaged in a moral violation) affects force judgment because of its effect on perceptions of the alternatives. To test this claim, we ran two multiple mediation analyses using the *mediation* package in R (Tingley et al., 2014), evaluating the degree to which unusualness and value judgments mediate the effect of condition on force judgements. In the first mediation analysis we designated unusualness judgments as the main mediator and value judgments as the alternative mediator. In the second mediation analyses we did the reverse, designating value judgments as the main mediator and unusualness judgments as the alternative. In both mediation analyses we included scenario as a covariate and confidence intervals were generated using 1000 bootstrapped samples. We found that, controlling for the mediating effect of value judgments, unusualness judgments significantly mediate the effect of condition on force judgment (Average mediation effect = 23.4%, 95% CI [13.61%, 33.1%]). However, controlling for the mediating effect of unusualness judgments, value judgments do not significantly mediate the effect of condition on force judgments (Average mediation effect = 5.65%, 95% CI [-8.73%, 20.03%]).

5.2 Discussion

Here we demonstrate, once again, that participants' force judgments are strongly dependent on their evaluations of possible alternative actions. Critically, consistent with Hypothesis 4, force judgments in this experiment are made in direct response to scenarios in which we manipulate whether or not agents engage in moral violations. The strong effect of evaluations of alternatives on force judgments, even in this context, provides compelling evidence against the claim that force judgments in response to moral transgressions are merely the result of evaluations of the actual action. Rather they suggest that, consistent with Phillips & Knobe's (2009) original claim and counterfactual relevance accounts more broadly, participants are more likely to view agents as acting freely when committing moral violations because they view these agents as having other options. Moreover, we build on this claim by providing evidence that the immorality of the actual action may make alternative actions appear more available because alternative actions appear more normal.

It is, perhaps, not surprising that alternative actions to moral transgressions are perceived as a better idea. We found, however, that unusualness judgments, but not value judgments, mediate the effect of condition on force judgments. This suggests that descriptive norms may be central to force judgments, even in highly prescriptive contexts like in response to moral transgressions.

Finally, in this study we show that evaluations of alternative actions drive force judgments even when possible alternative actions are not provided to the participant. In this experiment we merely ask participants to evaluate the unusualness and value of "doing something else." Therefore, alternative actions must be endogenously generated. This leaves open the possibility that, across participants, a vast set of possible alternative actions may have been envisioned. Nevertheless, we find that when making judgments about moral transgressions, participants tend to view the alternative actions as being more normal.

6. General Discussion

Counterfactual relevance accounts argue that the effect of morality on judgments of force and freedom result from the effect of morality on the relevance of counterfactual alternatives. We tested four predictions that arise from such accounts, and now want to consider how each fared in light of our findings.

6.1 Hypothesis 1: The effect of non-moral norm violations on force judgments.

While motivated moral reasoning accounts give moral violations a privileged influence on force judgments, counterfactual relevance accounts argue that moral norms are just one of many different norm violations that will impact the relevance of possible alternatives and subsequently affect the degree to which agents are viewed as forced. We tested this by asking participants to make force judgements in response to scenarios where we manipulated the degree to which actions and their alternatives adhered to prudential norms (Experiment 1) and social norms (Experiments 2 and 3). In Experiments 1 and 2, we found that even in these non-moral domains, force judgments were predicted by whether or not there was an available alternative action that was as or more normal than the actual action taken. While we did not find an effect of this manipulation in Experiment 3, we did find that across all three experiments the perceived normality of alternative actions significantly predicted the degree to which agents were viewed as forced. Put simply, the more participants perceived possible alternative actions as adhering to prudential or social norms, the more they viewed agents as acting freely when they chose not to do those alternative actions.

Taken in conjunction, this collection of results suggests that in the domain of prudential and social norms, the perceived normality of available counterfactual alternatives drives force judgments. While these findings are consistent with counterfactual relevance accounts, they are not straightforwardly predicted by motivated moral reasoning accounts, which claim that agents are viewed as having acted freely in order to hold them responsible for a moral violation or bad outcome. In fact, it is worth noting that in Experiments 1 and 3, the agents' actual actions are not norm violations at all—rather they are norm congruent (e.g. rejecting an item from the survival pack that would not be useful on a desert island in Experiment 1, or engaging in the same action most people do in Experiment 3). Given these features of our study design, it seems highly unlikely that participants' force judgments are motivated by a desire to blame, punish, or hold agents responsible for norm violations or producing bad outcomes.

6.2 Hypothesis 2: The effect of normality on force judgments depends on counterfactuals.

Counterfactual relevance accounts argue that the normality of the actual action affects force judgments primarily because it affects the degree to which possible alternative actions are seen as relevant. Therefore, we should find that shifting the perceived normality of the alternatives will lead to changes in force judgments, even when the actual action is held fixed. In Experiments 1 and 3 we find exactly this. As described above, in both experiments the agents choose a normative action, but are presented with alternatives of varying degrees of normality. And in both experiments, we find that as participants' perceptions of the normality of the alternatives increase, their perceptions that the agent acted freely also increase.

6.3 Hypothesis 3: The effect of both prescriptive and descriptive norms on force judgments.

Normality isn’t merely constrained to prescriptive normality (what is good, valuable, moral, or right) but also includes descriptive normality (what is likely or commonplace). Counterfactual relevance accounts argue that the relevance of possible alternatives is dictated by the alternatives’ prescriptive *and* descriptive normality relative to the actual action. Therefore, assessments of the degree to which alternatives are good or of the degree to which they are unusual should both affect force judgments. In all four experiments we find that the degree to which alternatives are viewed as a good idea or as unusual significantly predicts the degree to which agents are perceived as forced. Interestingly, the effect of unusualness persists even in Experiment 3, when we see no effect of value on force judgments. This suggests that descriptive norms can and do act independently on force judgments, even when we eliminate any effect of prescriptive normality.

6.4 Hypothesis 4: The effect of normality on force judgments in the moral domain.

As described above, counterfactual relevance accounts argue that both the prescriptive and descriptive normality of possible alternatives affects force judgments. This stands in contrast to motivated moral reasoning accounts, which argue that it is the morality of the action, in particular, that affects force judgments. According to counterfactual relevance accounts, then, we should expect that the same factors that explain force judgments *outside* of the moral domain, should explain force judgments *inside* the moral domain as well. In Experiment 4, we found that in cases where an agent engages in a morally wrong (e.g. a captain throwing his wife overboard to save his sinking ship) vs. morally neutral (e.g. the captain throws cargo overboard to save his sinking ship) action, the degree to which participants view the agent as forced is predicted by the degree to which they view possible alternatives as both a good idea *and* as unusual. In fact, we find that the effect of unusualness judgments on force judgments is significantly stronger than the effect of value judgments on force judgments. We also find that unusualness judgments mediate the effect of the moral valence of the actual action on force judgments when controlling for the mediating effect value judgments. However, the reverse is not true: When controlling for the mediating effects of unusualness judgments, we do not find that value judgements mediate the effect of moral valence of the actual action on force judgments.

6.5 Understanding the effect of normality on non-moral judgments

On the whole, our results indicate a clear pattern: Force judgments are driven by the normality of possible alternatives. The more alternative actions are viewed as normal, the more agents are perceived as acting freely. Counterfactual relevance accounts, however, argue that the normality of counterfactual alternatives is not special in its influence on force judgments. Rather, many different types of judgments seem to be influenced by the relevance of possible alternatives: Judgments of causation, intention, and doing versus allowing, all seem to depend on considerations of counterfactuals. Given our findings here, the effect of normality on all of these types of judgments may extend beyond whether an action or its alternatives violates a *moral* norm. Rather this should hold true in cases where prudential, social, or other types of norms are at issue. In light of this, we intend the present work to serve as a case study for the effect of normality on all of these types of judgments. Just as the degree to which alternatives adhere to this wide range of norms influences perceptions of force, so may they influence the degree to which agents are viewed as having been causal, having acted intentionally, and so on. In fact, this proposal finds some support in prior research which found that whether an agent is considered to have acted intentionally is influenced in the same way by violations of social/conventional norms (Uttich & Lombrozo, 2010) and moral norms (Knobe, 2003; Petit & Knobe, 2009). However, future research that systematically tests the effect of non-moral norm violations on these types of judgments remains necessary.

6.6 Why does descriptive normality matter for force judgments?

Our results also suggest that the descriptive normality of alternatives may be at least as important as the prescriptive normality. Why would this be the case? One possibility is that evaluations of the descriptive normality of alternatives may be influencing participants’ perceptions of the alternatives’ value. After all, actions that are taken by most people are often done so because they are the best choice. Likewise, morally wrong actions are much less commonplace than morally neutral or good ones. Therefore, participants may be inferring some kind of lower prescriptive value inherent in unusual actions, even in cases where we took great lengths to eliminate differences in prescriptive value.

While it is surely possible that individuals infer value from descriptively normal actions, there are several reasons why this is unlikely to explain our results. First, not only do we find that the perceived unusualness of alternatives predicts force judgments when there is no relationship between the perceived value of alternatives and force judgments (Experiment 3), but we find that even when the value of alternatives does predict force judgments, there is an effect of the unusualness of alternatives *over and above* the effect of value (Experiments 1, 2, and 4). Second, even in response to morally wrong vs. neutral actions, evaluations of unusualness mediate the effect of the moral valence of the action above and beyond the mediating effect of evaluations of value. These findings suggest that unusualness is not merely serving as a proxy for value in our study, rather it is having its own, independent effect on force judgments.

An alternative explanation for the surprising effect of descriptive normality on force judgments is that the normality of actions may influence participants’ *implicit* or *default* representations of how possible these counterfactual alternatives are. While the unusualness of an action is unlikely to influence participants’ explicit assessment of how possible that action was, there is good evidence that judgments of force rely on default rather than reflective judgments of possibility (Phillips & Cushman, 2017). In particular, these studies demonstrated that participants’ force judgments were better predicted by participants’ judgments of what it was possible for the agent to do when participants were put under time pressure, than when they were asked to reflect. In line with this proposal, there is a growing amount of evidence that default representations of possibility are closely linked to perceptions of what is moral or good, which fits well with our finding that an alternative option's prescriptive value predicts judgments of force (Phillips, Morris, Cushman, 2019).

The present work suggests that implicit or default representation of possibility may also be constrained by an action's descriptive normality. While additional research is still needed to confirm this possibility, this proposal does find some support from the developmental literature. Previous research found that not only do young children say that events that violate physical laws (e.g. eating lightening) are impossible, but events that violate prescriptive norms (e.g., stealing candy) and also descriptive norms (e.g., a boy wearing a dress) are impossible (Browne & Woolley, 2004; Kalish, 1998; Kushnir, Gopnik, Chernyak, Seiver, & Wellman, 2015; Shtulman, 2009; Shtulman & Carey, 2007). In fact, children claim that actions that violate both prescriptive and descriptive norms, just like those that violate physical laws, require magic (Phillips & Bloom, *under review*; Shtulman & Phillips, 2019). Our findings here build on this research by suggesting that default representations of possibility are not just shaped by prescriptive norms, but by descriptive norms as well. This remains an important area for future study.

6.7 Conclusions

The current study suggests that the degree to which alternative actions adhere to prescriptive or descriptive norms affects the degree to which agents are seen as having been forced to take the action they did. These findings fit best with counterfactual relevance accounts of the effect of morality, not just on force judgments but other types of non-moral judgments as well. They are difficult to explain by appealing to motivated moral reasoning, as many of the actions participants were judging in this study occurred outside the moral domain, and even when no norm violation or bad outcome occurred at all. More broadly, our results suggest a unifying role of normality and counterfactuals across many areas of high-level human cognition making this an important avenue for continued research.

Acknowledgments

Special thanks to Fiery Cushman for his ongoing help and support during the entire research process.

References

Alicke, M. (2008). Blaming badly. *Journal of cognition and culture*, *8*(1-2), 179-186.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670-696.

Adams, F., & Steadman, A. (2004). Intentional action in ordinary language: Core concept or pragmatic understanding?. *Analysis*, *64*(2), 173-181.

Bates, Douglas, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Fabian Scheipl, and Gabor Grothendieck. "Package 'lme4'." *Linear mixed-effects models using S4 classes. R package version* 1, no. 6 (2011).

Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition*, *167*, 25-37.

Browne, C. A., & Woolley, J. D. (2004). Preschoolers' magical explanations for violations of physical, social, and mental laws. *Journal of Cognition and Development*, *5*(2), 239-260.

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, 136, 30-37.

Clark, C. J., Baumeister, R. F., & Ditto, P. H. (2017). Making punishment palatable: Belief in free will alleviates punitive distress. *Consciousness and cognition*, *51*, 193-211.

Clark, C. J., Luguri, J. B., Ditto, P. H., Knobe, J., Shariff, A. F., & Baumeister, R. F. (2014). Free to punish: a motivated account of free will belief. *Journal of personality and social psychology*, *106*(4), 501-513.

Clark, C. J., Winegard, B. M., & Shariff, A. F. (2021). Motivated free will belief: The theory, new (preregistered) studies, and three meta-analyses. *Journal of Experimental Psychology: General, 150*(7), e22–e47.

Cushman, F., Knobe, J., & Sinnott-Armstrong, W. (2008). Moral appraisals affect doing/allowing judgments. *Cognition*, 108(1), 281-289.

Darley, J. M., & Shultz, T. R. (1990). Moral rules: their content and acquisition. *Annual Review of Psychology, 41,* 525–556

Everett, J. A. C., Clark, C. J., Meindl, P., Luguri, J. B., Earp, B. D., Graham, J., ... & Shariff, A. F. (2021). Political differences in free will belief are associated with differences in moralization. *Journal of Personality and Social Psychology*, *120*(2), 461.

Fillon, A. Lantian, A. Feldman, G., N'gbala, A. (2021). Exceptionality Effect in Agency: Exceptional Choices Attributed Higher Free Will Than Routine. *International Review of Social Psychology*.

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review, 128*(5), 936–975.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *The British Journal for the Philosophy of Science*, *66*(2), 413-457.

Harvey, J. H., Harris, B., & Barnes, R. D. (1975). Actor-observer differences in the perceptions of responsibility and freedom. *Journal of Personality and Social Psychology*, *32*(1), 22.

Hindriks, F. (2014), Normativity in Action: How to Explain the Knobe Effect and its Relatives. *Mind & Language*, *29*: 51-72.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587-612. Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.

Kalish, C. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development*, *69*(3), 706-720.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*(3), 190-194.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, 33(4), 315-329.

Knobe, J., & Szabó, Z. G. (2013). Modals with a taste of the deontic. *Semantics and Pragmatics*, *6*, 1-1.

Kominsky, J. F., & Phillips, J. (2019). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive science*, *43*(11), e12792.

Kominsky, J. F., Phillips, J., Gerstenberg, T., Lagnado, D., & Knobe, J. (2015). Causal superseding. *Cognition*, 137, 196-209.

Kushnir, T., Gopnik, A., Chernyak, N., Seiver, E., & Wellman, H. M. (2015). Developing intuitions about free will between ages four and six. *Cognition*, *138*, 79-101.

McCloy, R., & Byrne, R. M. (2000). Counterfactual thinking about controllable events. *Memory & Cognition*, *28*(6), 1071-1078.

Mele, A. R. (2003). Intentional action: Controversies, data, and core hypotheses. *Philosophical Psychology*, *16*(2), 325-340.

Malle, B. F., & Nelson, S. E. (2003). Judging mens rea: The tension between folk concepts and legal concepts of intentionality. *Behavioral Sciences & the Law*, *21*(5), 563-580.

Mandelkern, M., & Phillips, J. (2018). Sticky situations: 'Force' and quantifier domains. In *Semantics and Linguistic Theory*, *28*, 474-492).

Monroe, A. E., & Ysidron, D. W. (2021). Not so motivated after all? Three replication attempts and a theoretical challenge to a morally motivated belief in free will. *Journal of Experimental Psychology: General*, *150*(1), e1.

Nadelhoffer, T. (2004). Blame, Badness, and Intentional Action: A Reply to Knobe and Mendlow. *Journal of Theoretical and Philosophical Psychology, 24*(2), 259–269.

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

N'gbala, A., & Branscombe, N. R. (1995). Mental simulation and causal attribution: When simulating an event does not affect fault assignment. *Journal of Experimental Social Psychology*, *31*(2), 139-162.

Pettit, D., & Knobe, J. (2009). The pervasive impact of moral judgment. *Mind & language*, *24*(5), 586-604.

Phillips, J., & Bloom, P. (2017). Do children believe immoral events are impossible. *Under revision*.

Phillips, J., & Cushman, F. (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, *114*(18), 4649-4654.

Phillips, J., & Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychological Inquiry*, 20(1), 30-36.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65-94.

Phillips, J., Luguri, J. B., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30-42.

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*, *100*(2), 283-301.

Uttich, K., & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87-100.

Shtulman, A. (2009). The development of possibility judgment within and across domains. *Cognitive Development*, *24*(3), 293-309.

Shtulman, A., & Carey, S. (2007). Improbable or impossible? How children reason about the possibility of extraordinary events. *Child Development*, *78*(3), 1015-1032.

Shtulman, A., & Phillips, J. (2018). Differentiating "could" from "should": Developmental changes in modal cognition. *Journal of Experimental Child Psychology*, *165*, 161-182.

Tingley, D. Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis.

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, 119(2), 166-178.

Supplementary Information

Public Repository

Raw data, analysis scripts, and QSF files for running the experiment are publicly available on Harvard Dataverse at https://doi.org/10.7910/DVN/B8WAKM.

Experiment 1a Detailed Methods

In this study participants were told that the game show contestants actually participated in one of two different types of games. In the REJECT game, contestants were given a survival pack with three items and had to choose one item to reject from the survival pack, taking the other two items with them to the desert island. In the KEEP game, contestants were given a survival pack with three items and had to choose one item to *keep* from the survival pack. Only this one item would come to the desert island and the other two items would be left behind. Every participant completed 12 trials of each game, completing all trials of one game before playing the other game. The games were presented in random order across participants.

Within each game the survival pack contained either two items that would not be useful on a desert island, and one that would (Low-Low-High) or two items that would be useful on a desert island and one that would not (Low-High-High). The combination of game type (keep vs. reject) and survival pack contents (Low-Low-High vs. Low-High-High) created our two main conditions. The "Normal Alternative" condition consisted of two different trial types: when there was a Low-Low-High pack in the REJECT game, or a Low-High-High pack in the KEEP game. In the first case (LLH/REJECT), when the contestant rejects a low-value item from the pack, there is another low-value item (an equivalently normal alternative) he could have rejected instead. In the second case (LHH/KEEP), when the contestant keeps a high-value item from the pack, there is another high-value item (an equivalently normal alternative) he could have kept instead. The "Abnormal Alternative" condition consisted of the opposite combination of survival pack types and games: when there was a Low-High-High pack in the REJECT game, or a Low-Low-High pack in the KEEP game. In the first case (LHH/REJECT), when the contestant rejects a low-value item from the pack, there is no other alternative option that would be of equivalent normality (the only other items he could reject are high-value). In the second case (LLH/KEEP), when the contestant keeps the high-value item, there is no other alternative option that would be of equivalent normality (the only other items he could keep are low-value).

In addition to manipulating the game and contents of the pack, on each trial participants were also asked to respond to one of two different prompts. In HAD trials, participants were asked to rate how strongly they agreed with statements like, "Seems like [the contestant] had to reject the [rejected item]." or "I think [the contestant] had to keep the [kept item]." These trials were used to measure participants' force judgments. In DID trials, participants were asked to rate how strongly they agreed with statements like, "I think [the contestant] rejected the [rejected item]." or "Seems like [the contestant] kept the [kept item]." These trials were used to measure participants' understanding of what the contestant actually did. Across all trial-types, these statements were presented either with affirmative wording (e.g. "Seems like [the contestant] **had to** reject the [rejected item].") or negative wording (e.g. "Seems like [the contestant] **didn't have to** reject the [rejected item]."). This was done so that both agree and disagree responses were appropriate for all conditions, depending on whether the question was worded affirmatively or negatively.

Among the HAD trials, four were in the Normal Alternative condition (LLH/REJECT/Negative, LLH/REJECT/Affirmative, LHH/KEEP/Negative, LHH/KEEP/Affirmative) and four were in the Abnormal Alternative condition (LHH/REJECT/Negative, LHH/REJECT/Affirmative, LLH/KEEP/Negative, LLH/KEEP/Affirmative). There were also the same 8 trial types for the DID trials, however, for the present project, only data from HAD trials was analyzed. Responses to DID trials were excluded from the analyses (See SI Table 1).

Finally, across the whole study, 8 filler trials were included. In these trials, the contestant either rejected/kept an abnormal item or the participant responded to a statement that referenced an item that was not actually rejected or kept. Again, responses to these trials were excluded from the analyses.

Experiment 1b Attention Check Questions

After completing the main part of the experiment, participants responded to the following multiple choice attention check questions:

1. Which of the following items were not in the survival pack?
2. Which item did Joe (the contestant) decide to reject?

In addition they were asked to provide a written answer to the question: "Why did Joe need a survival pack?"

Experiment 3 Attention Check Questions

After completing the main part of the experiment, participants responded to the following multiple choice attention check questions:

1. Which of the following was *not* an option for [the agent] to use to [agent's activity as described in the scenario]?
2. Which of the following did [the agent] decide to use?
3. Why did [the agent] need to use [an object] to [agent's activity]?

SI Table 1.

Game	Condition	Survival Pack Contents	HAD/DID	Question Type (Affirmative/ Negative)	Data included in Experiment 1a
REJECT	Normal Alternative	Low-Low-High (Low value item rejected)	HAD	Affirmative	Yes
				Negative	Yes
			DID	Affirmative	No
				Negative	No
	Abnormal Alternative	Low-High-High (Low value item rejected)	HAD	Affirmative	Yes
				Negative	Yes
			DID	Affirmative	No
				Negative	No
KEEP	Normal Alternative	Low-High-High (High value item kept)	HAD	Affirmative	Yes
				Negative	Yes
			DID	Affirmative	No
				Negative	No
	Abnormal Alternative	Low-Low-High (High value item kept)	HAD	Affirmative	Yes
				Negative	Yes
			DID	Affirmative	No
				Negative	No

SI Table 1. Participants responded to one question in each of the trial types listed above. However, only the eight HAD trials were included in the analyses for the present project. Not shown are eight additional filler trials in which either the abnormal object was rejected/kept or participants were asked about an incorrect object.