Judgment and Decision Making, Vol. 14, No. 2, March 2019, pp. 120–129

A reason-based explanation for moral dumbfounding

Matthew L. Stanley* Siyuan Yin† Walter Sinnott-Armstrong‡

Abstract

The moral dumbfounding phenomenon for harmless taboo violations is often cited as a critical piece of empirical evidence motivating anti-rationalist models of moral judgment and decision-making. Moral dumbfounding purportedly occurs when an individual remains obstinately and steadfastly committed to a moral judgment or decision even after admitting inability to provide reasons and arguments to support it (Haidt, 2001). Early empirical support for the moral dumbfounding phenomenon led some philosophers and psychologists to suggest that affective reactions and intuitions, in contrast with reasons or reasoning, are the predominant drivers of moral judgments and decisions. We investigate an alternative reason-based explanation for moral dumbfounding: that putatively harmless taboo violations are judged to be morally wrong because of the high perceived likelihood that the agents could have caused harm, even though they did not cause harm in actuality. Our results indicate that judgments about the likelihood of causing harm consistently and strongly predicted moral wrongness judgments. Critically, a manipulation drawing attention to harms that could have occurred (but did not actually occur) systematically increased the severity of moral wrongness judgments. Thus, many participants were sensitive to at least one reason — the likelihood of harm — in making their moral judgments about these kinds of taboo violations. We discuss the implications of these findings for rationalist and anti-rationalist models of moral judgment and decision-making.

Keywords: moral psychology, social-intuitionist model, reasoning, disgust, risk, harm

1 Introduction

Some contemporary moral psychologists have developed influential models of moral judgment and decision-making that denigrate the causal role played by reasons and reasoning. Instead, affect and intuition are posited to be predominant causal forces responsible for bringing about moral judgments and decisions, especially when they are made in private (Greene & Haidt, 2002; Haidt, 2001, 2007, 2012; Haidt & Bjorklund, 2008; Prinz, 2006, 2007). According to these models, appeals to reasons and reasoning are primarily utilized in a post-hoc manner to support those pre-existing judgments and decisions formed from affective reactions and intuitions. As Haidt (2001) puts it, "moral reasoning does not cause moral judgment; rather, moral reasoning is usually a post hoc construction, generated after a judgment has been reached" (p. 814). Haidt goes on to claim that affective reactions and moral intuitions tend to drive moral reasoning "just as surely as a dog wags its tail" (p. 830). Those who emphasize the pre-eminence of affect and intuition over reasons and reasoning in producing moral judgments are commonly referred to as anti-rationalists.

What empirical evidence provides support for this anti-rationalist position? Among the most commonly cited evidence is *moral dumbfounding* (Haidt, 2001, 2007, 2012; Prinz, 2006, 2007). Moral dumbfounding purportedly occurs when a person makes a moral judgment in a particular situation, admits to being unable to adequately defend that judgment or decision with reasons and arguments, but still remains obstinately and steadfastly committed to that initial judgment (Haidt, 2001, 2007). The experiments investigating moral dumbfounding conducted by Haidt and colleagues utilize vignettes depicting putatively harmless but shocking taboo violations performed by unknown others, such as this one (Haidt et al., 2000):

> Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love?

The first two authors contributed equally to this work.

*Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Levine Science Research Center, Rm. B254, Duke University, Durham, NC 27708-0743. E-mail: matthew.stanley@duke.edu.

†Department of Philosophy, Duke Institute for Brain Science, Duke University.

‡Department of Philosophy, Kenan Institute for Ethics, Duke Institute for Brain Sciences, Duke University.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

Not only did most participants condemn Julie and Mark's behavior, but when pressed to justify their condemnation post-hoc, their initial justifications tended to appeal to harm-related features of incest in general, such as psychological distress. This was the case even though it was explicitly stated in the vignette that no one was harmed. When the experimenter rejected these harm-related justifications for condemnation, most participants still insisted that Julie and Mark's behavior was wrong. This continued insistence that a behavior is wrong while admitting that there is no adequate justification for the judgment is the moral dumbfounding phenomenon.

The results from Haidt et al. (2000) may seem particularly compelling and surprising because of the substantial percentage of participants who do not reverse their initial judgments, despite their inability to overcome countervailing arguments and reasons provided by the experimenter. Before attempting to justify their judgments, 20% of the participants initially stated that Julie and Mark's actions were OK, meaning that 80% of participants initially stated that their behavior was not OK. After failing to provide satisfactory justification as to why the behavior was not OK, only 32% of participants ended up claiming that the behavior was OK.

The case of Julie and Mark is not the only one that elicits a moral dumbfounding effect. Similar effects have been found for other taboo violations involving homosexuality, unusual masturbation, and eating a pet dog that just died in a car accident (Haidt, 2001, 2007).[1] The experimenters attempted to design each vignette such that (1) a shocking taboo is violated by at least one other person, (2) nobody in the vignette experiences any harm, and (3) most participants would explicitly admit that nobody was hurt by the behavior. In doing so, the experimenters attempted to ensure that participants could not appeal to actual harms to justify their condemnation.[2]

Why do people seem so resistant to changing their judgments in these cases despite their failures to provide acceptable justifications? Anecdotally, Haidt et al. (2000) reported that many participants who became morally dumbfounded referred to their own emotional reactions as final attempts to justify or explain their judgments. That is, as a last resort, many would end up proclaiming that the behaviors are just disgusting or revolting. This finding was corroborated by Haidt and Hersh (2001), who reported that when participants attempted to justify why purportedly harmless taboo violations are wrong, many cited their own affective reactions of shock, disgust, and discomfort (Haidt & Hersh, 2001). Providing more direct evidence, Haidt and Hersh (2001) also found that self-reported negative affect was related to wrongness judgments for another set of shocking taboo violations involving homosexuality, unusual masturbation, and consensual incest.[3] Haidt (2001, 2007, 2012) has argued that this moral dumbfounding phenomenon is the product of strong, automatic affective reactions. Participants were filled with disgust and revulsion upon reading about Mark and Julie's behavior, and they could not ignore or override such a strong, affective reaction — even when they failed to provide adequate reasons to justify their condemnation.

This phenomenon has captured the attention of philosophers, psychologists, and the general public, and it has influenced popular theories of moral judgment and decision-making. Nevertheless, the available evidence for moral dumbfounding is subject to challenge. For example, participants might have assumed that providing certain principles (e.g., incest is wrong or, more generally, disgusting or repugnant acts are wrong; cf. Kass, 1997) as reasons for moral condemnation would not have been satisfactory for the experimenter. That is, participants might have assumed that the experimenter wanted something *more* than just the stating of the principle. But stating such a principle does offer a reason for moral condemnation. It is also possible that participants interpreted the question about whether Julie and Mark's behavior was "OK" as a question concerning social norms or laws that the participants themselves would not necessarily have endorsed. In this way, participants might have indicated that the behavior was "not OK" as a way of reporting what is commonly thought by other people, even if the participants did not personally believe that there were good reasons to condemn the behavior.

Empirical evidence has also been garnered to challenge Haidt's (2001, 2007) explanation for the moral dumbfounding phenomenon. Although Haidt and colleagues attempted to carefully devise the vignettes to be void of harm, it does not follow that participants actually believed the behaviors to be harmless (Gray, Schein, & Ward, 2014; Royzman et al., 2015). People have a tendency to comprehend and interpret fictional content by relating it to their real-world knowledge, often finding it difficult to grasp and accept information within a fictional world that contradicts what they know or believe to be true in the real world (Ferguson & Sanford, 2008; Ferguson, Scheepers & Sanford, 2010; Royzman et al., 2015). Using the original Julie and Mark vignette devised by Haidt et al. (2000), Royzman et al. (2015) reported that many

[1]Note that the percentage of participants experiencing moral dumbfounding and refusing to change their judgments after running out of reasons and arguments does vary considerably depending upon the vignette. Across eight different vignettes (and political orientation), Haidt and Hersh (2001) found that anywhere between 7% and 60% of participants experienced moral dumbfounding and refused to change their judgments.

[2]It is important to note that the supposed evidence for the moral dumbfounding phenomenon comes only from shocking taboo violations that are purportedly harmless. This is a relatively specific, narrow kind of moral violation. It is unclear whether anybody becomes dumbfounded when presented with more commonplace moral violations, such as those involving harm, unfairness, dishonesty, or justice (Hofmann et al., 2014), that are not necessarily expected to be held without question.

[3]Note, however, that this evidence does not permit conclusions to be drawn about negative affect *causing* moral wrongness judgments.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

participants explicitly commented on their difficulties imagining how the siblings' relationship would remain unaffected in the aftermath of the action. In fact, most participants refused to accept that Julie and Mark's decision to have sex had no negative impact on their relationship. Interestingly, these participants still went on to exhibit all the classic signs of a morally dumbfounded state, as reported by Haidt et al. (2000), including: confusion, admission of lack of reasons, and yet refusal to reconsider their moral judgments. In this way, participants may frequently reject the harm-negating provisos contained within these vignettes, ultimately believing that harms actually did occur, but they still exhibit the outwardly observable characteristics of being dumbfounded (see also Gutierrez & Giner-Sorolla, 2007).

Here we offer a distinct, novel explanation for supposed moral dumbfounding: that the perceived *risk* of causing harm is associated with moral condemnation. Of course, there are some circumstances in which actual consequences track moral judgments: people tend to think that an agent who got unlucky and caused harm deserves more blame than an otherwise identical agent who got lucky and did not cause harm (Levy, 2016). But in many cases, even if individuals cannot appeal to actual harms to justify their condemnation of an act, they can still reasonably appeal to higher-order beliefs about the *likelihood* of the act causing harm. To illustrate this point, consider the following analogy: driving drunk is considered morally objectionable, even when the drunk driver actually makes it home safely, because of the risk of causing harm to others (or himself). Similarly, in the case of Julie and Mark, although the reader is informed that no harm actually befell Julie or Mark (or anyone else), their sexual acts risked causing harm to one another or to other people (e.g., family members who found out about their behavior). As Haidt notes, some participants did cite "the dangers of inbreeding" and the possibility that "Julie and Mark will be hurt, perhaps emotionally" (2001, p. 814). Participants were then reminded that "no harm [actually] befell them" (2001, p. 814), but this reply by the experimenter does not deny that their behavior was risky or that their behavior could have caused harm. In the story, Julie and Mark did not know beforehand that they would be lucky and that no harm would befall anyone. As in the case of the drunk driver who made it home safely, people might still condemn Julie and Mark's behavior because each created a risk of causing harm to others.

Many people might agree that such taboo violations are actually harmless but still morally wrong because they create danger or risk of harms, such as psychological distress or damaged relationships. Risk of harm then serves as a reasonable justification for condemnation even in the absence of any actual harm. Then, they can cite reasons for condemning the acts without rejecting the assurances in these vignettes that no actual harm occurred.

Utilizing vignettes depicting a variety of putatively harmless taboo violations, we first provide evidence for a strong positive relationship between the perceived likelihood of the agents causing harm and moral wrongness judgments (Experiments 1, 2a, and 2b). We then demonstrate that the perceived likelihood of causing harm, emotional responses to the violations, and moral wrongness judgments are all strongly and consistently correlated with each other for a variety of putatively harmless taboo violations (Experiments 2a and 2b). This suggests that prior moral dumbfounding research identifying an association between affective reactions and moral judgments may be confounded by the perceived risk of the behavior. Finally, we explicitly manipulate the salience of possible harms that could have occurred (but did not actually occur) to show that even brief reflection on potential harms increases the severity of moral wrongness judgments (Experiment 3). Taking all this evidence together, it cannot be concluded that the emotional response to the violation is *the* driver of moral judgments for putatively harmless taboo violations. Instead, our results offer a reason-based alternative explanation for moral dumbfounding.

2 Experiment 1

The purpose of Experiment 1 is to provide correlational evidence for the relationship between the perceived risk of causing harm and moral wrongness judgments.

2.1 Materials and Method

Participants. 250 individuals participated for a small payment in this study on Amazon's Mechanical Turk (AMT). Participant recruitment was restricted to individuals in the United States who had at least 50 previously accepted HITs and a prior approval rating above 90%. Twelve participants failed the attention check at the end or did not answer all questions in the experiment, so data were analyzed with the remaining 238 individuals (M_{age} = 35 years, SD = 10, $range_{age}$ = [18, 70], 170 females).

Materials. Two vignettes describing putatively harmless taboo violations involving incest and cannibalism, respectively, were used in this study. These two vignettes were adapted from Haidt et al. (2000), but they were modified to reduce the skewness of moral wrongness judgments. The *Julie and Mark* vignette describes first cousins who have sex on a vacation. The *Jennifer* vignette describes an individual who cooks and eats a piece of human flesh. See Appendix A for exact materials.

Procedure. This study consisted of a single self-paced session. First, one of the two randomly chosen vignettes was

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

provided to participants. After reading the vignette, participants were asked to indicate how likely it is that there could have been harmful consequences on a 7-pt scale from 1 (*very unlikely*) to 7 (*very likely*). Then, participants rated the moral wrongness of the behavior on a 7-pt scale from 1 (*not at all morally wrong*) to 7 (*very morally wrong*). Participants then repeated this procedure for the second vignette.

After answering all questions for both vignettes, participants were asked the following attention check: "Do you feel that you paid attention, avoided distractions, and took the survey seriously?" They responded by selecting one of the following: (1) no, I was distracted; (2) no, I had trouble paying attention; (3) no, I didn't not take the study seriously; (4) no, something else affected my participation negatively; or (5) yes. Participants were ensured that their responses would not affect their payment or their eligibility for future studies. Only those participants who selected (5) were included in the analyses.

2.2 Results and Discussion

To test our hypothesis that the perceived likelihood of causing harm predicts judgments of moral wrongness, we computed correlations between these two variables for each vignette separately. The perceived likelihood that harm could have occurred was positively correlated with moral wrongness judgments for both vignettes (*Julie and Mark*: $r(236) = 0.64$, $p < 0.001$, 95% CI = [0.56, 0.71]; *Jennifer*: $r(235) = 0.42$, $p < 0.001$, 95% CI = [0.31, 0.52]). For both vignettes, the greater the perceived risk of causing harm, the more severe the moral wrongness judgments.

3 Experiment 2a

Experiment 1 provides some initial evidence that that the perceived risk of harm strongly predicts moral wrongness judgments. The primary purpose of Experiment 2a is to investigate the extent to which the perceived likelihood of harm, emotional responses to the violation, and moral judgments are all correlated with each other. Other researchers have argued that affective responses (and disgust in particular) drive moral wrongness judgments for putatively harmless taboo violations (e.g., Haidt, 2001, 2012; Prinz, 2007). Although we grant that affective responses, such as feelings of disgust, likely influence moral judgments for some individuals in some cases (Landy & Goodwin, 2015), moral dumbfounding research identifying an association between affective reactions and moral judgments could be confounded by the perceived risk of the behavior. The results of Experiment 2a indicate that the perceived likelihood of harm, emotional responses to the violation, and moral wrongness judgments are all strongly and consistently correlated with each other. These results suggest that that the perceived risk of causing harm could be a confound in previous research that has identified a relationship between emotion and moral judgments for putatively harmless taboo violations.

3.1 Materials and Method

Participants. 230 individuals voluntarily participated in this study on Amazon's Mechanical Turk (AMT). Participant recruitment was restricted to individuals in the United States who had at least 50 previously accepted HITs and a prior approval rating above 90%. 25 participants failed the attention check at the end or did not answer all questions, so data were analyzed with the remaining 205 individuals (M_{age} = 34 years, SD = 10, $range_{age}$ = [18, 69], 84 females). Those who participated in the previous experiment were prevented from accessing the HIT.

Materials. The same two vignettes used in Experiment 1 was also used in Experiment 2a (see Appendix A).

Procedure. This study consisted of a single self-paced session. First, one of the two randomly chosen vignettes was shown to participants. After reading the vignette, participants made the following ratings in a randomized order: the likelihood of harmful consequences (1 = *very unlikely*, 7 = *very likely*); disgust (1 = *not at all disgusted*, 7 = *very disgusted*); anger (1 = *not at all angry*, 7 = *very angry*); sadness (1 = *not at all sad*, 7 = *very sad*); fear (1 = *not at all afraid*, 7 = *very afraid*); and distress (1 = *not at all distressed*, 7 = *very distressed*). All questions about emotions were specifically about how the participant felt while reading the scenario. Our primary interest is the relationship between of feelings of disgust, the perceived risk of harm, and moral wrongness judgments, but we included measures of anger, sadness, fear, and distress for two reasons. First, we intend to determine whether disgust is in fact the particular emotion most clearly associated with moral wrongness judgments for these putatively harmless taboo violations, as Haidt and others suggest (Haidt, 2001, 2012). We intentionally selected other emotions that might be associated with moral wrongness judgments. Second, the inclusion of these additional emotion measures helps to conceal the aims of our study. After making these ratings of risk and emotion, participants rated the moral wrongness of the behavior on a 7-pt scale from 1 (*not at all morally wrong*) to 7 (*very morally wrong*). Participants then repeated this procedure for the second vignette.

After answering all questions for both vignettes, participants were asked the same attention check question and excluded for reporting failures and distractions, as in the previous experiment. Upon completion, participants were monetarily compensated for their time.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

TABLE 1: Means, Standard Deviations, and Pearson Correlation Coefficients in Experiment 2a Split by Vignette.

Variables	M	SD	1	2	3	4	5	6
Julie and Mark Vignette								
Moral Judg.	4.59	2.09	.					
Disgust	4.47	2.09	.80	.				
Anger	2.70	1.84	.55	.51	.			
Sadness	2.80	1.88	.49	.49	.69	.		
Fear	2.24	1.73	.28	.28	.57	.60	.	
Distress	3.32	2.03	.59	.65	.68	.72	.55	.
Perceived risk	4.26	1.84	.61	.66	.45	.46	.26	.53
Jennifer Vignette								
Moral Judg.	6.02	1.58	.					
Disgust	6.04	1.59	.71	.				
Anger	4.32	2.15	.49	.45	.			
Sadness	3.51	2.11	.29	.26	.58	.		
Fear	3.39	2.14	.22	.21	.42	.54	.	
Distress	4.79	1.95	.48	.50	.61	.58	.56	.
Perceived risk	5.17	1.77	.57	.55	.41	.32	.28	.47

Note. $N = 205$.

3.2 Results and Discussion

Means, standard deviations, and bivariate correlations between all variables are available in Table 1. First, to replicate our results from Experiment 1, we computed zero-order correlations between the perceived likelihood of harm having occurred and moral wrongness judgments for each vignette taken separately. The perceived likelihood of harm having occurred was significantly and positively correlated with moral wrongness judgments for both vignettes (*Julie and Mark*: $r(203) = 0.61$, $p < 0.001$, 95% CI = [.52, .69]; *Jennifer*: $r(203) = 0.57$, $p < 0.001$, 95% CI = [.47, .66]). To address our primary question of interest, for each vignette, we computed a series of zero-order correlations, finding that feelings of disgust, the perceived likelihood of causing harm, and moral judgments were all strongly and significantly interrelated for both vignettes (all ps < .001). These correlation coefficients ranged from .61 to .80 in the Julie and Mark vignette, and from .55 to .71 in the Jennifer vignette (Table 1).

4 Experiment 2b

Having found that feelings of disgust, the perceived likelihood of causing harm, and moral judgments are all strongly and significantly inter-related for both vignettes in Experiment 2a, the purpose of Experiment 2b is to investigate whether this same pattern of results is also present using a different set of vignettes describing other taboo violations.

4.1 Materials and Method

Participants. 205 individuals voluntarily participated in this study on Amazon's Mechanical Turk (AMT) with monetary compensation. Participant recruitment was restricted to individuals in the United States who had at least 50 previously accepted HITs and a prior approval rating above 90%. Five participants failed the attention check at the end or did not answer all questions, so data were analyzed with the remaining 200 individuals (M_{age} = 36 years, SD = 11, $range_{age}$ = [19, 74], 72 females). Those who participated in the previous experiments were prevented from accessing the HIT.

Materials. Four new vignettes were adapted from Parkinson et al. (2011) and used in Experiment 2b (see Appendix B).

Procedure. The procedure and exclusion criteria in Experiment 2b are the same as those in Experiment 2a. The only difference between Experiments 2a and 2b is the particular vignettes used.

4.2 Results and Discussion

Means, standard deviations, and bivariate correlations between all variables are available in Table 2. First, we computed correlations between the perceived likelihood of causing harm and moral wrongness judgments for each vignette taken separately. The perceived likelihood of causing harm was significantly and positively correlated with moral wrongness judgments for all four vignettes (*Phil*: $r(198) = .54$, $p < .001$, 95% CI = [0.43, 0.63]; *Tim*: $r(198) = .58$, $p < .001$, 95% CI = [0.48, 0.66]; *James and Holly*: $r(198) = .63$, $p < .001$, 95% CI = [0.53, 0.70]; *Ann and Bob*: $r(198) = .60$, $p < .001$, 95% CI = [0.50, 0.68]). The more likely it is that the behavior could have caused harm, the more severe the moral wrongness judgments in four vignettes. To address our primary question of interest, for each vignette, we computed a series of zero-order correlations, finding that feelings of disgust, the perceived likelihood of causing harm, and moral judgments were all strongly and significantly inter-related for all vignettes (all ps < .001). These correlation coefficients ranged from .42 to .62 in the Phil vignette, from .38 to .58 in the Tim vignette, from .59 to .71 in the James and Holly vignette, and from .59 to .74 in the Ann and Bob vignette.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

TABLE 2: Means, Standard Deviations, and Pearson Correlation Coefficients in Experiment 2b Split by Vignette.

Variables	M	SD	1	2	3	4	5	6
Phil Vignette								
Moral Judg.	5.48	2.00	.					
Disgust	5.97	1.53	.62	.				
Anger	3.31	2.12	.45	.43	.			
Sadness	3.44	2.15	.43	.37	.66	.		
Fear	2.40	1.87	.26	.24	.67	.66	.	
Distress	4.32	2.07	.45	.54	.62	.68	.55	.
Perceived Risk	4.80	2.01	.54	.42	.41	.37	.34	.45
Tim Vignette								
Moral Judg.	5.56	1.89	.					
Disgust	6.29	1.24	.49	.				
Anger	3.98	2.21	.51	.43	.			
Sadness	3.57	2.23	.42	.30	.64	.		
Fear	2.70	2.00	.31	.22	.57	.63	.	
Distress	4.53	2.07	.51	.45	.70	.61	.54	.
Perceived Risk	4.93	1.81	.58	.38	.45	.30	.31	.49
James and Holly Vignette								
Moral Judg.	5.25	2.05	.					
Disgust	5.20	2.04	.71	.				
Anger	2.95	2.06	.49	.48	.			
Sadness	3.14	2.09	.46	.42	.66	.		
Fear	2.24	1.75	.34	.31	.63	.63	.	
Distress	3.76	2.10	.55	.62	.63	.67	.53	.
Perceived Risk	4.93	1.85	.63	.59	.40	.42	.31	.49
Ann and Bob Vignette								
Moral Judg.	4.74	2.03	.					
Disgust	4.80	2.00	.74	.				
Anger	2.60	1.90	.49	.55	.			
Sadness	2.78	1.92	.45	.50	.72	.		
Fear	2.05	1.67	.32	.35	.70	.66	.	
Distress	3.37	2.06	.51	.61	.66	.64	.58	.
Perceived Risk	4.10	1.95	.60	.59	.52	.47	.34	.48

Note. $N = 200$

5 Experiment 3

In Experiment 3, we make salient potential harms that could have occurred (but did not in actuality) to show that even brief reflection on risks increases the severity of moral wrongness judgments. In doing so, we provide evidence for the positive conclusion that experimentally manipulating the salience of possible harms that could have occurred systematically influences moral wrongness judgments. Experiment 3 thus serves to disambiguate the causal direction of influence driving the correlations obtained in our previous experiments.

5.1 Materials and Method

Participants. 660 individuals voluntarily participated in this study on Amazon's Mechanical Turk (AMT) in return for a small payment. Participant recruitment was restricted to individuals in the United States who had at least 50 previously accepted HITs and a prior approval rating above 90%. 24 participants failed the attention check at the end or did not answer all questions, so data were analyzed with the remaining 636 individuals (M_{age} = 36 years, SD = 12, $range_{age}$ = [19, 74], 294 females). Those who participated in previous experiments were prevented from accessing the HIT.

Materials. The same two vignettes used in Experiments 1and 2a were also used in Experiment 3 (see Appendix A).

Procedure. This study consisted of a single self-paced session. Participants were randomly assigned to one of the two vignettes in a between-subjects fashion. After reading the vignette, participants were randomly assigned to the experimental condition or the control condition. In the experimental condition, participants read about three ways in which the behavior could have caused harm even though it did not in actuality (see Appendix C). In the control condition, participants read about three neutral ways in which the event could have gone differently even though it did not in actuality (Appendix C). Regardless of the condition to which participants were assigned, they all rated the moral wrongness of the behavior on a 7-pt scale from 1 (*not at all morally wrong*) to 7 (*very morally wrong*).

After answering all questions for both vignettes, participants were asked the same attention check question as in previous experiments, and we excluded participants who reported being distracted, having trouble paying attention, failing to avoid distractions, and not taking the survey seriously.

5.2 Results and Discussion

We tested the hypothesis that, when possible but non-actual harms are made salient, participants will judge the behaviors to be more morally wrong relative to a matched control condition. For the *Julie and Mark* vignette, after participants

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

read about three ways in which the behavior could have caused harm (M = 5.04, SD = 1.93), they rated the behavior as more morally wrong relative to those who read about three neutral ways in which things could have been different (M = 4.57, SD = 2.20; n = 329, $t(325.03)$ = 2.04, p = 0.04, 95% CI = [0.02, 0.91], Cohen's d = .23). Similarly, for the *Jennifer* vignette, after participants read about three ways in which the behavior could have caused harm (M = 6.46, SD = 0.97), they rated the behavior as more morally wrong relative to those who read about three neutral ways in which things could have been different (M = 6.11, SD = 1.41; n = 307, $t(282.51)$ = 2.55, p = 0.01, 95% CI = [0.08, 0.62], Cohen's d = .29). These results indicate that making salient ways in which harm could have occurred (but did not occur in actuality) increases the severity moral wrongness judgments relative to a matched control condition.

6 General Discussion

The moral dumbfounding phenomenon has attracted considerable interest and has played a central role in recent models of moral judgment and decision-making, both directly (Cushman et al., 2010; Haidt, 2001, 2012; Prinz, 2006, 2007) and indirectly (Crockett, 2013; Ditto, Liu & Wojcik, 2012; Greene, 2014). Moral dumbfounding, by definition, occurs when people remain obstinately and steadfastly committed to their moral judgments even after becoming aware of their inability to provide reasons and arguments to support those judgments (Haidt, 2001). Based on results from empirical investigations into the moral dumbfounding phenomenon, philosophers (e.g., Prinz, 2006, 2007) and psychologists (e.g., Haidt, 2001, 2012) alike have then argued that affective reactions and intuitions are really the primary drivers of moral judgments. The purportedly harmless taboo violations used to investigate moral dumbfounding effects are thought to elicit strong feelings of disgust and revulsion, which in turn, drive moral condemnation. They argue that appeals to reasons and reasoning are then predominantly utilized in a post-hoc manner to support those pre-existing judgments and decisions formed from intuitions and affective reactions.

To investigate the moral dumbfounding phenomenon, Haidt and colleagues developed a set of vignettes in which at least one person commits a taboo violation, but no harm actually befalls anyone (Haidt et al., 2000; Haidt & Hersh, 2001). In this way, they take away the ability of participants to appeal to *actual* harms as reasons for moral condemnation. Still, the agents in the vignettes did not know that they would not end up causing harm. There was a non-trivial chance that they could have caused harm to someone, even though the actors in the vignettes got lucky and did not end up harming anyone.

Using the same kinds of purportedly harmless taboo violations as Haidt and colleagues, we examined the relationship between moral judgments and a different explanatory variable — the risk of agents causing harm (even when they actually cause no harm). Overall, we found that judgments about the likelihood of causing harm consistently and strongly predicted moral wrongness judgments. Critically, manipulating the salience of potential harms systematically changed the severity of moral wrongness judgments. In contrast, existing evidence for moral dumbfounding does not include manipulations of affect, so this research does not show that affective reactions *cause* moral condemnation. In any case, whether or not affect is also a cause, many participants were sensitive to at least one reason — the risk of causing harm — in making their moral judgments about these kinds of taboo violations.

Although we do not deny that affect and intuition play a role in forming and changing some moral judgments for some people in some circumstances, our results indicate that moral dumbfounding does not provide the requisite evidence for concluding that reasons and reasoning play only a minor or inconsequential role in bringing about moral judgments and decisions — even for purportedly harmless taboo violations. Appealing to the risk of causing harm that did not occur in actuality offers a reason-based justification of moral wrongness judgments. This alternative variable significantly predicts moral wrongness judgments and is perfectly consistent with the harm-negating provisos explicitly stated in the vignettes.

Consider again the case of the drunk driver who made it home safely: people tend to condemn the drunk driver who made it home safely because of the high probability of harming others (or himself). Similarly, in cases of incest or bestiality — even when no harm actually befalls anyone — there is still a clear risk of causing harm to someone, and the agents do not know that they will not cause any harm. Many people find it difficult to articulate such reasons about risks, so we should not infer from the premise that someone is unable to immediately provide reasons for her moral judgments in one particular situation to the conclusion that she lacks reasons altogether (May, 2018; Saltzstein & Kasachkoff, 2004). It remains possible that reasons guided the moral judgments of participants in previous investigations of moral dumbfounding, but that participants struggled to consciously articulate their reasons at that moment, perhaps because of social pressures in the interview setting that encouraged them to be cooperative and non-combative (Royzman et al., 2015). So, many participants who supposedly experienced dumbfounding in prior studies could actually have been making their judgments based on the perceived risk of causing harm. Additionally, taboo violations like incest and bestiality are often difficult for people to think about and talk about, and many people might not have ever consciously considered explicit justifications for why such

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

behaviors might be morally wrong. Because of this, it may be difficult for participants to explicitly articulate why such behaviors are morally wrong when asked by an experimenter.

The identified relationship between perceived risk of harm and moral judgments in our studies also raises a normative question. We might condemn the drunk driver who makes it home safely because of the high probability that he could have harmed others (or himself). But we frequently engage in other behaviors that carry certain risks that we do not condemn. For example, when we drive (sober) to the grocery store, we are imposing a potential risk of harm on some possible would-be-victim that would not have been imposed if we had walked to the grocery store instead. Why do we condemn the drunk who drives home but not the sober driver making a trip to the grocery store? In other words, where do we draw the line between what is moral or immoral when people engage in risky behavior? We do not offer an answer to this normative question, but some philosophers have discussed this issue at length (e.g., Kumar, 2015), and the issue is somewhat analogous to the definition of negligence in tort law (e.g., Shavell, 2004).

If existing empirical investigations into the moral dumbfounding phenomenon do not provide the requisite evidence for anti-rationalist models of moral judgment and decision-making, is there any reason to hold an anti-rationalist position? There is evidence that some people tend to use reasons and arguments in a motivated, biased way to support their moral judgments and decisions. For example, Stanley et al. (2018) presented participants with several different two-option moral dilemmas. Participants rarely changed their minds after evaluating reasons against their initial decisions, and they tended to evaluate new reasons in a biased way to lend support to their initial decisions. This biased evaluation ultimately made participants more confident in their initial decisions. Similar effects have also been found for certain applied ethical issues, such as the death penalty (Lord et al., 1979) and drone strikes on military targets overseas (Stanley et al., 2019). Moreover, using dilemmas indexing the permissibility of sacrificing one innocent person to save a greater number of people, Uhlmann and colleagues (2009) found that people flexibly and selectively appeal to moral principles that support their own judgments in a way that is consistent with their political leanings. Importantly, these studies do not directly show that reasons play a minor or inconsequential role in forming moral judgments and decisions. They show only that many people do sometimes evaluate reasons in a biased, post-hoc manner to lend support to their pre-existing judgments and decisions.[4] Taking all this evidence together, biases in reasoning and reason-evaluation do seem to play some role in moral judgment and decision-making, which could support some weak anti-rationalist positions.

Nevertheless, none of this evidence lends support to the strong position that people typically have no reasons when they make their moral judgments and decisions. On the contrary, there is evidence that reasons and reasoning do play a role in forming and changing at least some moral judgments and decisions in a less-biased manner (May, 2018). For example, whether individuals agree with a particular moral principle predicts their judgments across a range of moral dilemmas (Lombrozo, 2009), and being reminded that they agree with a particular moral principle can make many people change their moral judgments (Horne et al., 2015). Future work will further examine the conditions under which reasons are more likely to actively change moral judgments or serve a post-hoc confirmatory function.

[4]There is an important distinction to be made here between biases in reason evaluation and having no reasons whatsoever. See Royzman et al. (2015) for a similar point.

References

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. *The Oxford handbook of moral psychology*, 47–71.

Ditto, P. H., Liu, B., & Wojcik, S. P. (2012). Is anything sacred anymore?. *Psychological Inquiry*, *23*(2), 155–161.

Ferguson, H. J., & Sanford, A. J. (2008). Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, *58*(3), 609–626.

Ferguson, H. J., Scheepers, C., & Sanford, A. J. (2010). Expectations in counterfactual and theory of mind reasoning. *Language and Cognitive Processes*, *25*(3), 297–346.

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600.

Greene, J. D. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D., & Haidt, J. (2002). How (and where) does moral judgment work?. *Trends in Cognitive Sciences*, *6*(12), 517–523.

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. *Emotion*, *7*(4), 853–868.

Haidt, J. (2001). The emotional dog and its rational tale: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834.

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, *316*(5827), 998–1002.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Vintage.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press

Haidt, J., Bjorklund, F., & Murphy, S. (2000). Moral dumbfounding: When intuition finds no reason. *Lund Psychological Reports, 2*, 1–23.

Haidt, J., & Hersh, M. A. (2001). Sexual morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology, 31*, 191–221.

Hofmann, W., Wisneski, D. C., Brandt, M. J., & Skitka, L. J. (2014). Morality in everyday life. *Science*, *345*, 1340–1343.

Horne, Z., Powell, D., & Hummel, J. (2015). A single counterexample leads to moral belief revision. *Cognitive Science, 39*, 1950–1960.

Kass, L. (1997). The Wisdom of Repugnance. *The New Republic, 216*(22), June 2, 17–26.

Kumar, R. (2015). Risking and wrongdoing. *Philosophy and Public Affairs, 43*, 27–51.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science*, *10*(4), 518–536.

Levy, N. (2016). Dissolving the puzzle of resultant moral luck. *Review of Philosophy and Psychology*, *7*(1), 127–139.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*(2), 273–286.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109.

May, J. (2018). *Regard for reason in the moral mind*. Oxford University Press.

Parkinson, C., Sinnott-Armstrong, W., Koralus, P. E., Mendelovici, A., McGeer, V., & Wheatley, T. (2011). Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust. *Journal of Cognitive Neuroscience*, *23*(10), 3162–3180.

Prinz, J. (2006). The emotional basis of moral judgments. *Philosophical Explorations, 9*(1), 29–43.

Prinz, J. (2007). *The emotional construction of morals*. Oxford University Press.

Royzman, E. B., Kim, K., & Leeman, R. F. (2015). The curious tale of Julie and Mark: Unraveling the moral dumbfounding effect. *Judgment and Decision Making*, *10*(4), 296–313.

Saltzstein, H. D., & Kasachkoff, T. (2004). Haidt's moral intuitionist theory: A psychological and philosophical critique. *Review of General Psychology, 8*, 273–282.

Shavell, S. (2004) *Foundations of economic analysis of law*. Cambridge, MA: Belknap Press of Harvard University Press.

Stanley, M. L., Dougherty, A. M., Yang, B. W., Henne, P., & De Brigard, F. (2018). Reasons probably won't change your mind: The role of reasons in revising moral decisions. *Journal of Experimental Psychology: General, 147*, 962–987.

Stanley, M. L., Henne, P., Yang, B. W., & De Brigard, F. (2019). Resistance to position change, motivated reasoning, and polarization. *Political Behavior*. https://doi.org/10.1007/s11109-019-09526-z.

Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*, 479–491.

Appendix A

Julie and Mark Vignette. Julie and Mark, who are first cousins, are traveling together in France. They are both on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy it, but they decide not to do it again. They keep that night as a special secret between them, which makes them feel even closer to each other.

Jennifer Vignette. Jennifer works in a medical school pathology lab as a research assistant. The lab prepares human cadavers that are used to teach medical students about anatomy. The cadavers come from people who had donated their body to science for research. One night Jennifer is leaving the lab when she sees a body that is going to be discarded the next day. Jennifer was a vegetarian, for moral reasons. She thought it was wrong to kill animals for food. But then, when she saw a body about to be cremated, she thought it was irrational to waste perfectly edible meat. So, she cut off a piece of flesh, and took it home and cooked it. The person had died recently of a heart attack, and she cooked the meat thoroughly, so there was no risk of disease.

Appendix B

Phil Vignette. Phil is visiting his 67-year old grandmother. As she is baking, he comes up behind her and kisses her passionately. They are both aroused and start to rub up against each other until they climax. This encounter never creates problems for either of them, and neither experiences any harm.

Tim Vignette. Tim is hiking in a secluded forest one afternoon when he discovers a freshly dead male coyote. He has not seen anyone of the trail all day, so he has anal intercourse with the coyote body, using a condom. Afterwards, he continues his hike and never suffers any negative effects.

James and Holly Vignette. James and Holly are brother and sister. After they both graduate from college, they share an apartment in a large building. When nobody else is around, they sometimes touch each other's genitals passionately. This activity never creates any problems for either of them. Neither of them ever suffers any harm from doing this.

Ann and Bob Vignette. Ann and Bob are adult siblings. When they are alone, they decide to kiss each other on the mouth passionately, using their tongues. They do this only once, and neither of them ever suffers any negative effects.

Appendix C

Below are three ways in which the behavior of Julie and Mark could have caused harm:

1. Their birth control methods could have failed, and Julie could have gotten pregnant.
2. Making love could have actually damaged their relationship.
3. At some later time, their friends and family could have found out that they had sex.

Below are three neutral ways in which the event could have gone differently in the Julie and Mark vignette:

1. They could have been traveling in Switzerland.
2. They could have been staying in a cabin in the mountains.
3. They could have been on a trip during spring break.

Below are three ways in which the behavior of Jennifer could have caused harm:

1. Someone could have seen that Jennifer was carrying the piece of human flesh out of the lab.
2. The dead person could have been carrying some unknown disease that Jennifer was unaware of.
3. Someone could have found out that Jennifer ate the piece of flesh at some later time.

Below are three neutral ways in which the event could have gone differently in the Jennifer vignette:

1. Jennifer could have been working in a nursing school pathology lab.
2. Jennifer could have left the lab at 6:10 PM.
3. Jennifer could have been living in a townhouse instead of an apartment.

https://doi.org/10.1017/S1930297500003351 Published online by Cambridge University Press