COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2018) 1–31

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12598

Intuitive Probabilities and the Limitation of Moral Imagination[†]

Arseny A. Ryazanov,[a] Jonathan Knutzen,[b] Samuel C. Rickless,[b]
Nicholas J. S. Christenfeld,[a] Dana Kay Nelkin[b]

[a]*Department of Psychology, University of California San Diego*
[b]*Department of Philosophy, University of California San Diego*

Received 5 May 2017; received in revised form 12 January 2018; accepted 16 January 2018

Abstract

There is a vast literature that seeks to uncover features underlying moral judgment by eliciting reactions to hypothetical scenarios such as trolley problems. These thought experiments assume that participants accept the outcomes stipulated in the scenarios. Across seven studies (N = 968), we demonstrate that intuition overrides stipulated outcomes even when participants are explicitly told that an action will result in a particular outcome. Participants instead substitute their own estimates of the probability of outcomes for stipulated outcomes, and these probability estimates in turn influence moral judgments. Our findings demonstrate that intuitive likelihoods are one critical factor in moral judgment, one that is not suspended even in moral dilemmas that explicitly stipulate outcomes. Features thought to underlie moral reasoning, such as intention, may operate, in part, by affecting the intuitive likelihood of outcomes, and, problematically, moral differences between scenarios may be confounded with non-moral intuitive probabilities.

Keywords: Moral judgment; Morality; Probability; Intuition; Trolley problem

1. Introduction

There is a large and important literature in moral theory centered around eliciting participants' reactions to scenarios of various kinds and drawing conclusions about their judgments. For example, systematic psychological investigation focused on the widely studied Trolley Problems has uncovered a variety of features of scenarios that may account for participants' moral judgments, including whether the agent is described as

Correspondence should be sent to Arseny A. Ryazanov, Department of Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0109. E-mail: aryazano@ucsd.edu

[†]The order of the senior authors is arbitrary.

causing harm to another through physical contact (e.g., Cushman, Young, & Hauser, 2006), whether information about motives is provided (e.g., Nichols & Knobe, 2007), whether harm is described as being intentional (e.g., Cushman et al., 2006, 2006; Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Mikhail, 2000; Moore, Clark, & Kane, 2008; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006), and many other factors.

In arriving at these conclusions, it is typically simply assumed that participants have accepted the claims made about features of the scenarios, and in particular claims about what outcomes will occur if agents act in certain ways (though there are some notable exceptions such as Royzman and Baron's [2002] and Greene et al.'s [2009] studies, discussed below). In a series of studies, we set out to test this assumption. Perhaps when participants are told that, in a given scenario, if an agent acts in a particular way a certain outcome will ensue, participants do not accept this claim. Instead, perhaps they only accept that a certain outcome *might* ensue. If this hypothesis is correct, then participants' judgments could be affected by a factor that is not built into the scenario at all and is in fact at odds with the scenario as described. Further, if participants are in fact substituting their own judgments about the probabilities of outcomes for scenario stipulations, this should lead us to rethink the way in which participants' reactions to scenarios are tested and what conclusions can be drawn. Our experiments focused mainly on scenarios eliciting moral judgments, but we believe that the relevance of these findings potentially extends to the use of scenarios intended to elicit all sorts of judgments, including moral, prudential, and linguistic, among others.

To see what is at stake, consider Philippa Foot's (1967) classic Trolley Problem, which has been the centerpiece of an entire literature in moral theory. A host of variations of the case has been used as a test of intuitions, and the results have been thought by many to provide strong support for a particular kind of moral theory over its main rival. One variant, developed by Judith Jarvis Thomson (1976), consists of two scenarios. In one, call it "Side Track," five people are tied to a trolley track and will be killed unless a bystander pulls a lever that switches the trolley onto a side track. One person is tied to the side track and will be killed if the bystander pulls the lever. In another, call it "Footbridge," five people are tied to a trolley track and will be killed unless a bystander pushes a large man off a bridge above the track, in which case his body will stop the trolley. If the bystander pushes the large man off the bridge, he will be killed, but the five will be saved. A large majority of participants judges that it is morally permissible to turn the trolley in Side Track, but morally impermissible to push the large man in Footbridge (see, e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001). But in each case, if the bystander turns the trolley or pushes the large man, five will be saved and one will die. Thus, if the only morally relevant factor in determining permissibility were consequences—as Consequentialist moral theories have it—then it should be permissible *both* to turn the trolley and to push the large man. Are we simply inconsistent and mistaken in our judgments? Or is there another morally relevant factor (or factors) that can explain in a principled way why turning the trolley is permissible and pushing the large man is not? Some moral theorists have offered the following explanation: In Footbridge, unlike in

Side Track, the one is *used* without his consent, if the bystander saves the five (see, e.g., McIntyre, 2004). It is plausible that we have a right not to be used against our will, even when the consequences would be better overall if we were. This fact, it is claimed, explains the moral difference between the two cases, removing the need to attribute inconsistency in most participants' reactions.

The trolley cases continue to be used by philosophers as "intuition pumps," and in classrooms as well as in journals intuitive reactions to the cases are employed in important arguments for non-consequentialism. Intuitions about cases are typically used as part of a methodology known as "reflective equilibrium" (see Rawls, 1971). On this approach, one tries to reach a kind of equilibrium between plausible general principles, on the one hand, and intuitions about particular cases, on the other. This might require rejecting some intuitions in favor of others, if intuitions are ultimately inconsistent with each other or with plausible principles. While the approach is subtle and does not rely only on intuitions about cases, it is nevertheless true that intuitions play a powerful role in reasoning to a particular moral theory.

As these cases have caught the attention of experimental psychologists in recent years, the burgeoning research program that tests participants' reactions to the cases in systematic and creative ways has yielded intriguing results. One such result is that in cases in which philosophers had initially thought that only a single factor (such as whether or not the victim was used against his will in the plan to save others) explains our differential judgments, the cases also vary on other dimensions that appear to play an explanatory role, such as whether or not there is physical contact in the case (e.g., Cushman et al., 2006). If it turns out that participants' moral judgments are affected by what many philosophers have taken to be morally irrelevant features of cases, then this is very important information for moral theorists to have. For example, if on reflection participants would disavow the idea that directness of physical contact makes a moral difference, then moral theorists cannot claim to have successfully isolated a single morally salient factor that is both explaining and justifying participants' initial moral judgments. Further, this is also important information for psychologists to have in working to identify the mechanisms operative in moral reasoning, and for anyone interested in human behavior in a wide variety of contexts, from decisions in wartime to emergency rooms to legislatures.

We set out to study various versions of moral principles featuring in ordinary moral thought, isolating morally salient features of situations in turn. But some of us, in teaching ethics, had been struck that even after setting out hypothetical cases clearly and repeatedly, students often explicitly commented that they just were not going to accept the stipulated features in the case. This experience led us to try to test whether participants in experiments also resist accepting stipulated features of the cases and, if so, whether this affects their judgments of permissibility.

We are not the first to raise this set of issues. Some researchers have expressed concern about resistance to stipulated features (Bennis, Medin, & Bartels, 2010; Christensen & Gomila, 2012; Greene et al., 2009). Some have tried to control for participants' substitution of their own intuitive probability judgments. For example, Royzman and Baron

(2002), studying whether participants make different moral judgments when agents harm victims "directly" and "indirectly," eliminated from consideration those participants who they found had judged differentially on probability of outcomes across "direct" and "indirect" scenarios. But they did so along with several other factors and did not test what contribution, if any, differential probability judgments make in generating differential moral judgments.

Greene et al. (2009, p. 365) also tried to control in two ways for what they call "unconscious realism," that is, "a tendency to unconsciously replace a moral dilemma's unrealistic assumptions with more realistic ones." First, they told participants the scenarios were "unrealistic" and asked them to suspend disbelief about them, eliminating responses from participants who reported being unable to do so. Second, they asked participants who had previously evaluated the moral acceptability of protagonists performing actions in the various scenarios how probable it was that each protagonist's action would be (a) as described in the dilemma (e.g., five lives saved at the cost of one), (b) worse than this, or (c) better than this. They found that ratings of higher probability that the outcome would be worse than as described were correlated with lower moral acceptability ratings. As Greene et al. (2009) pointed out, this raises the question of whether participants' substitution of their own probability estimates of outcomes affects their moral judgments. But, as they also noted, participants' probability estimates might have been offered as post hoc rationalizations for their prior moral judgments, and the evidence was at most for a correlation and not causation. So, while they assumed the existence of unconscious realism in an attempt to control for it in investigating other aspects of the scenarios that affect moral judgments, and while their results were suggestive of the hypothesis that the effects of unconscious realism "may be real," they did not themselves offer studies that test the effect of probability estimates alone on moral judgments. Rather, they urged others to think about controlling for this possibly real effect in future studies.

Finally, Kortenkamp and Moore (2014) focused directly on assessing probability as a factor in moral judgment, but they did not explore whether participants substitute their own probability judgments when responding to scenarios in which outcomes were stipulated as certain. They asked for participants' probability judgments only in connection with scenarios in which outcomes of protagonists' actions were already described as *uncertain* to happen (e.g., "x might happen" and "you don't know whether" [p. 380]). They did, however, find that participants' responses concerning moral rightness or wrongness differed between scenarios in which outcomes were stipulated as certain and scenarios in which outcomes were stipulated as uncertain. Yet, interestingly, when they probed for probabilities in the uncertain scenarios, they did not find a significant correlation between probability judgments and moral judgments (although they did find a small effect when it came to expected value judgments and moral judgments). Nevertheless, taken as a whole, these studies, while not showing that people substitute probabilities for stipulated certainty, do show that moral judgments are impacted by people thinking about outcomes as probabilistic.

In the studies described below, we directly put to the test the hypothesis that participants substitute their own estimates of the probability of outcomes for those stipulated in

the scenarios, and that their doing so affects their moral judgments about cases. Unlike in Kortenkamp and Moore's (2014) study, we presented scenarios in which the outcomes are stipulated and in which the protagonists know that particular outcomes will occur, to test whether participants are in fact substituting their own judgment for the stipulated features.

And unlike Kortenkamp and Moore (2014), Royzman and Baron (2002), and Greene et al. (2009), we presented pairs of scenarios that varied only in outcomes that differed in their antecedent probabilities (such as whether someone would die from a train running over their foot or their neck). Our approach also differs in that we distinguished among a variety of specific probabilities, including both the probability that the one will die and that the five will live if the agent acts and also both the probability that the one will live and that the five will die if the agent does *not* act. We tested whether participants substitute their own estimates of the probability of any one of these outcomes, and whether any one of these probability estimates could affect participants' moral judgments.

Study 1 has four parts. Although they were initially part of the same experiment, for ease of exposition we present the first two as Studies 1a and 1b. In Study 1a, we presented participants with a Trolley Problem that was designed to vary on whether or not the harm to the single person was a means to save the five, or simply a foreseeable result of saving them. We asked participants to report not only how permissible they thought the action was in the two cases, but also how likely they thought the action was to actually save the five and kill the one. Study 1b also presented participants with pairs of cases designed to isolate changes in the perceived likelihood of outcomes by keeping fixed additional factors such as whether or not the harm was intended as a means. In this study, we limited the changes to antecedent probabilities that the one would die if the protagonist took action (e.g., turning the trolley). Studies 1c and 1d provide evidence that the participants' lack of acceptance of certain outcomes was not a result of the pragmatics of asking them to respond to the likelihood of the event, nor a particular result of participants' use of a slider.

To control for the possibility that participants' responses concerning the likelihood of outcomes might influence their moral assessments, in Study 2 we presented the same scenarios as in Studies 1a and 1b and asked participants only for their moral assessments of the scenarios, without first asking about the likelihood of outcomes.

Studies 3 and 4 parallel the probability-isolating scenarios from Study 1b. Instead of varying the antecedent probability of the one dying if the protagonist took action, we instead varied the antecedent probability of the five being saved if the protagonist took action.

For the first four studies, we employed the same sort of moral judgment scale used by Cushman et al. (2006), namely, a Likert scale, which runs from 1 (impermissible) to 7 (permissible). Because many philosophers assume that the concept of moral permissibility is binary rather than scalar, in Study 5, we presented participants with one of the scenarios from Study 4, using only a Yes/No question concerning whether the protagonist's action would be morally permissible.

In Study 6, we extended the range of probability estimates assessed. People may bring intuitive probabilities to bear not only on the likelihood of the action resulting in the death of the single person and the saving of the five, but also on the risk to each group in the case of inaction. To investigate how different probability estimates concerning each of four different outcomes might affect moral judgment, we presented participants with four scenario sets using a within-subjects design and asked them for four different probability judgments (two concerning how probable outcomes are if the protagonist acts and two concerning how probable outcomes are if the protagonist does not act), as well as eliciting their moral judgments on a Likert scale.

Finally, in Study 7, we investigated whether scenarios designed to capture moral distinctions in quite abstract terms are nevertheless perceived as varying in perceived probabilities.

2. Study 1a

2.1. Procedure

One hundred and twenty-one participants located within the United States were recruited as participants via Amazon's Mechanical Turk (M_{age} = 35.6, SD = 13.2; 62.8% female). We used a pair of Trolley Cases, adapted from Thomson (1976), written to capture the intended/foreseen distinction, forms of which often appear in the moral judgment literature (e.g., Cushman et al., 2006). In one member of the pair, the death of one individual is a foreseen consequence of diverting a trolley that will otherwise run over five individuals. In the other, the individual's body is instrumental to saving the five: pushing him in front of the trolley stops it from hitting the five. In both scenarios, the death of the one and saving of the five are stipulated as outcomes of the protagonist's action. However, pushing a person in front of a trolley may seem more likely to kill the one, as well as less likely to save the five, than diverting the trolley. (See Supplementary Materials for scenarios.)

Participants were also presented with three other scenario sets that will be discussed subsequently as Study 1b. Scenarios were presented in a randomized order. After reading each scenario, participants were asked to respond to how likely they thought the death of the single individual would be if the protagonist decided to perform the action, using a percentage scale (e.g., *If Sam decides to divert the trolley in order to save the five, how likely is it that the lone individual will die?; 0–100%)*. They were also asked to estimate a second likelihood: the likelihood that the five would be saved should the protagonist decide to perform the action (e.g., *If Sam decides to divert the trolley in order to save the five, how likely is it that the five will be saved?; 0–100%*). Next, participants were asked to rate the permissibility of the action on a 7-point scale (*How permissible would it be for Sam to divert the trolley in order to save the five?; 1 = Impermissible to 7 = Permissible*). Participants also provided demographic information.

2.2. Results

The pair of scenarios was analyzed for within-subject differences in perceived likelihoods and permissibility using paired-samples *t* tests. Participants reported that the one was significantly more likely to die when his death would result from being pushed onto the tracks ($M = 92.9\%$, $SD = 15.8$), compared to when it would result from the trolley being diverted into him ($M = 89.2\%$, $SD = 20.3$), $t(120) = 2.83$, $p = .005$, $d = .20$ (see Table 1). Participants also reported perceiving differences in the likelihood of the five being saved, $t(120) = 3.60$, $p < .001$, $d = .36$, with diverting the trolley being more likely to save them than pushing a man in its way (mean push = 83.1%, $SD = 24.7$; mean divert = 91.2%, $SD = 19.5$). Differences in these perceived likelihoods were paralleled by differences in permissibility: action in the scenario perceived as more likely to result in the death of the one individual and less likely to save the five was also rated as significantly less permissible, $t(120) = 7.20$, $p < .001$, $d = .75$ (mean push = 2.96, $SD = 1.95$; mean divert = 4.46, $SD = 2.07$). In neither case did participants take the death of the one, stipulated as the outcome of the protagonist's action, to be 100% likely, nor did participants think that the protagonist's action was 100% likely to save the five, an outcome that was also explicitly stipulated.

A bootstrap mediation analysis with 10,000 resamples, using the mediation package in R (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014) revealed that differences in the perceived likelihood of the one dying did not mediate permissibility differences between the two scenarios (indirect effect $ab = .04$, $p = .13$, 95% CI [−.01, .12]; direct effect $c' = 1.45$, $p < .001$, CI [1.0, 1.9]; 3% mediated), nor did differences in the perceived likelihood of the five being saved mediate permissibility differences between the two scenarios (indirect effect $ab = .05$, $p = .31$, 95% CI [−.05, .16]; direct effect $c' = 1.45$, $p < .001$, 95% CI [.32, .65]; 3% mediated), failing to clearly demonstrate a relationship between permissibility and outcome likelihood, despite both significantly varying between the two scenarios.

2.3. Discussion

The two scenarios differ along the intended/foreseen distinction, but participants also appear to register a difference in the perceived likelihood of the one dying and the five

Table 1
Rated likelihoods and permissibility of scenario set intended to vary along the foreseen/intended distinction (Study 1a)

Version	Less Likely (divert trolley)	More Likely (push man)	Sig. Diff.
Rated likelihood of the one dying (%)	89.2	92.9	**
Rated likelihood of the five being saved (%)	91.2	83.1	***
Permissibility	4.46	2.96	***

Note. **$p < .01$, ***$p < .001$.

being saved. Why this difference in perception? One possibility is that the difference between the diversion of the trolley and the use of the bulk of the man as an obstacle to block the trolley's path is itself associated with a difference in the perceived likelihood of the outcome occurring, independently of the protagonist's intentions. Another is that people judge an intended harm as generally more likely to occur, since they think the protagonist would likely take extra steps to try to bring it about, whereas the protagonist might take steps to reduce the likelihood of its occurrence in the case of merely foreseen harm. Given that in one of the scenarios, the death of the one person was more likely to occur and that it was less likely for that death to be accompanied by the saving of five, it is reasonable, especially given other differences, that the agent's conduct was judged less morally permissible. We do not, however, find evidence for the perceived likelihoods mediating the permissibility judgments. It may be that the scales, especially an unfamiliar one on moral permissibility, may be too crude to capture participants' intuitions when so many factors are varying between scenarios.

In any case, two things are clear. First, people do not accept outcomes as they are stipulated in the moral scenarios. Second, scenarios written to vary on only one dimension can easily and inadvertently vary also in the perceived likelihood of the specified harm and benefit occurring. Since it is morally relevant how likely harms and benefits are to occur, it would not be surprising if judgments of this sort have some effect on participants' moral verdicts. For the next study, we explored whether participants' perceived likelihoods affect their permissibility judgments in sets of scenarios that are designed not to differ on any other potentially morally relevant dimension.

3. Study 1b

3.1. Procedure

As noted above, the participants of Study 1 were presented with three additional scenario pairs in which a protagonist is faced with an action that will save five individuals but kill one other. The members of each additional scenario-pair were parallel except for the way in which the single individual would die if the protagonist performed the action. The death of the one individual and the survival of the five were explicitly stipulated as the outcomes of action in every case. In one case, the mode of death, for example, the trolley severing his neck, had a plausibly higher perceived likelihood of being fatal than the mode of death in the other, for example, the trolley severing his foot. The other two scenario sets involved (a) pushing a man off a ledge, causing him to fall either 10 feet or 150 feet to his death in order to reach five to save them, and (b) pushing either small rocks or a large boulder off a ledge onto a person below in order to reach five in time to save them. (See Supplementary Materials for scenarios.) If participants substitute their perceived likelihoods of the one individual's death for the scenarios' clear assertion that the individual will die in each case if the agent acts, then they will rate the odds of his death as different between the two cases. This in turn might affect their assessment of the permissibility of the respective actions.

After reading each scenario, participants responded to how likely they thought the death of the single individual would be if the protagonist decided to perform the action, as well as to how likely they thought the saving of the five would be if the action was performed. They also rated the permissibility of the action on a 7-point Likert scale (1 = Impermissible to 7 = Permissible).

3.2. Results

We explored the contribution of altering the perceived likelihood of outcomes on permissibility using a repeated measures ANOVA, with scenario set and condition specified as fixed factors. First, we examined whether the predicted permissibility differences emerged from having manipulated the perceived likelihood that the one would die, despite stipulating that the death was certain. The ANOVA yielded a significant permissibility difference, $F(1, 120) = 21.3$, $p < .001$, partial $\eta^2 = .15$ (mean less likely = 4.68, $SD = 1.97$; mean more likely = 4.27, $SD = 2.00$; see Table 2). There was also a main effect difference between scenario sets on permissibility, $F(2, 240) = 8.48$, $p < .001$, partial $\eta^2 = .07$ (mean set 1 = 4.56, $SD = 2.08$; mean set 2 = 4.23, $SD = 1.97$; mean set 3 = 4.64, $SD = 1.92$), but there was no significant interaction between scenario set and condition, $F(2, 240) = 2.03$, $p = .13$, partial $\eta^2 = .017$, suggesting that all three scenario sets varied equally in permissibility.

Next we examined whether perceived likelihoods of the one dying varied within each scenario pair. A repeated measures ANOVA revealed that they did, $F(1, 120) = 47.1$, $p < .001$, partial $\eta^2 = .28$ (mean low = 79.0, $SD = 27.7$, mean high = 90.9, $SD = 18.2$). They also varied between scenario sets, $F(2, 240) = 4.18$, $p = .016$, partial $\eta^2 = .068$, (mean set 1 = 87.2, $SD = 22.5$; mean set 2 = 84.4, $SD = 25.2$; mean set 3 = 83.3, $SD = 24.6$). There was no interaction between scenario set and perceived likelihood of the one dying, $F(2, 240) = 1.98$, $p = .14$, partial $\eta^2 = .02$, suggesting that perceived likelihoods varied similarly across sets of scenarios.

We also explored whether our scenarios varied the perceived likelihood of the five being saved by the action. Our manipulation did not influence the perceived likelihood that the five would be saved, $F(1, 120) = .823$, $p = .37$, partial $\eta^2 = .007$ (mean low = 87.0, $SD = 21.8$; mean high = 87.6, $SD = 21.0$), though there was a significantly different perceived likelihood that the five would be saved between sets of scenarios, $F(2, 240) = 12.5$, $p < .001$, partial $\eta^2 = .09$ (mean set 1 = 91.7, $SD = 18.9$, mean set 2 = 85.3, $SD = 21.9$; mean set 3 = 84.9, $SD = 22.5$). There was no interaction between condition and scenario, $F(2, 240) = .98$, $p = .38$, $\eta^2 = .008$.

Next, a bootstrap mediation analysis with 10,000 resamples (Tingley et al., 2014) revealed that differences in the perceived likelihood of the one dying partially mediated the permissibility differences between pairs of scenarios (mediated effect $ab' = .22$, $p < .001$, 95% CI [.15, .30]; direct effect $c' = .19$, $p = .02$, 95% CI [.03, .36]; 53% mediated). These mediation models specified random intercepts and slopes for subjects.

Table 2
Rated likelihoods and permissibility across three scenarios sets that vary along perceived likelihood of the one dying despite death being stipulated (Study 1b)

	Scenario		
Version	Less Likely	More Likely	Sig. Diff.
Rated likelihood of the one dying (%)	79.0	90.9	***
Rated likelihood of the five being saved (%)	87.0	87.6	ns
Permissibility	4.68	4.27	***

Note. ***$p < .001$.

3.3. Discussion

Despite outcomes being explicitly stipulated, participants reported a less than 100% likelihood of their occurrence. They also reported divergent likelihoods of the one dying within each of the three pairs of scenarios. Furthermore, an increased perceived likelihood of the one dying as a result of the actions corresponded, quite reasonably, with a judgment of a lower degree of permissibility for the action relative to the other member of its pair: Judgments of the likelihood of the action resulting in the one person's death mediated, partially, judgments of the permissibility of the action. Given that the scenarios were designed to differ morally only in the intuitive likelihood of the action actually resulting in harm to the one, the finding that the mediation is only partial most likely reflects the intrinsic noisiness of participants' estimations of probabilities and permissibility.

Next, we explored two factors that might be thought to cause participants to report uncertainty (or lower than 100% probability) about the outcomes even if they really accepted the certain stipulation. One is the pragmatics of asking likelihood questions about a stipulated outcome, which might suggest that we want an answer less than 100%. The other is that our scale featured a slider that ranged from 0% to 100%, and so any error in reporting certainty would have to be on the low side, which would produce a mean of less than 100%. It is worth noting that such effects cannot readily account for the observed differences between our paired scenarios. If participants are just picking some likelihood arbitrarily less than 100% to make the question seem reasonable, or are just reporting 100% with some (bounded-at-100) error, then these effects should be equal in the two versions of the scenarios, and clearly that is not the case; participants are responding lawfully to perceived differences in outcomes. Nonetheless, we explored whether these two factors contribute to the less-than-certain reports in a scenario with a high perceived likelihood of the stipulated outcome actually coming to pass.

4. Study 1c and 1d

We used the scenario that had been rated as having an outcome closest to 100% from Study 1b (the trolley severs the neck of the one if the switch is thrown, rated as 94.5%

likely to result in the death of the one). If the difference from 100% is just due to random error, then when participants enter their response to the likelihood question freely, rather than using a slider capped at 100%, the error ought to be roughly symmetrical around 100%, and the average should be certainty. This is the approach of Study 1c. Study 1d examined the pragmatics issue by making the ratings not just of the one certain thing, but of a variety of factors whose perceived likelihoods reasonably varied, and included one whose likelihood was 100%. The addition of these other questions before the critical question about the stipulated certainty of the one person dying if the action is taken should reduce any suggestion that we want them to respond with a probability less than 100%.

4.1. Procedure 1c

Participants read the sever-the-head scenario and were asked, *If Rachel decides to divert the trolley in order to save the five, how likely is it that the lone individual will die (e.g., the likelihood of a flipped coin coming up heads is 50%)*? Participants entered any numeric value they wanted into a text box that appeared below the question.

4.2. Results 1c

Sixty-two participants (M_{age} = 34.6, SD = 10.3, 74.2% female) rated the action as having an 89.5% (SD = 24.7) likelihood of killing the lone individual, which significantly differed from 100%, $t(62) = 3.32$, $p = .002$, $d = .43$. Thus, even when entering a percentage value, with no slider-induced norm, participants still did not accept that the stated outcome would certainly happen.

4.3. Procedure 1d

Participants read the same scenario and answered four other likelihood questions before the critical one. These questions were chosen to elicit responses that varied, justifying the inquiry about likelihood, and also to include a question to which the participants would answer 100%, to normalize participants reporting 100% on the critical question, if that is what they felt the answer really was. Participants were asked how likely a fair six-sided die was to come up less than four, and how likely the die was to come up less than seven. They also were asked how likely it is that Rachel knows one of the five individuals on the track ahead, and how likely it is that she knows the lone person on the track. Finally, participants were asked the critical question: *If Rachel decides to divert the trolley in order to save the five, how likely is it that the lone individual will die?*

4.4. Results 1d

Sixty-four people (M_{age} = 24.5 SD = 12.3; 60.9% female) participated in the study. We included what we thought would be a simple math problem about the likelihood

of a regular six-sided die coming up less than seven, so that participants would see that we had at least one question where 100% was a reasonable answer. However, our math problem turned out to be insufficiently easy, and 16 (of 64) people did not respond with 100% on the likelihood of the die showing less than seven. To be conservative, we examined the ratings of the likelihood of the diverted trolley killing the lone individual for the participants who had reported 100% for the die roll (though the data are no different for all subjects). These participants on average reported that likelihood as 92.9% (SD = 21.6), again significantly different from 100%, $t(47) = 2.26$, $p = .029$, $d = .38$.

4.5. Discussion

Even when not using a slider that caps responses at 100%, and in the context of questions where asking about likelihood was quite ordinary and 100% was a reasonable answer, participants nonetheless did not accept that the action would certainly result in the stipulated death of the one. There was no indication of either the use of a slider or a potentially surprising presentation of questions concerning likelihood contributing to the original effect. Having conservatively selected the scenario rated as most likely to result in death for this test, it is likely that this pattern of findings generalizes to the other scenarios tested. This suggests that rated likelihoods are not the result of a norm dictated by the use of a slider or by the question being asked in isolation, but that participants are indeed not accepting stipulated outcomes, not only in the less likely versions of our scenarios, but in the more likely ones as well.

It is, however, possible that the difference in estimates of the permissibility of the action across each pair depended on participants being sensitized to the perceived likelihood of the act being fatal. That is, being asked to rate the likelihood of the action's fatality could have directed participants' attention to perceived likelihoods and thereby influenced permissibility judgments. The next experiment tested whether the permissibility difference we observed would persist if participants were not asked to rate the likelihood of action outcomes, but just read various scenarios that included stipulations of outcomes (as before). Would participants, in the absence of questions regarding likelihood, nevertheless find action in the scenarios previously found to be more likely to kill the one less permissible?

5. Study 2

5.1. Procedure

One hundred and nineteen participants located in the United States were recruited via Mechanical Turk (M_{age} = 34.4, SD = 11.1; 60.5% female). We repeated the procedure from Study 1, although with both questions regarding likelihoods removed, leaving a single permissibility question for each scenario.

5.2. Results

Even without questions regarding likelihood, action in the scenarios in which harm to the one had been previously judged to be more likely was rated as less permissible. In the scenario set from Study 1a, causing the death of the one by throwing a switch was judged more permissible ($M = 4.29$, $SD = 2.12$) than pushing him into the way of the trolley ($M = 2.96$, $SD = 1.90$), $t(118) = 7.48$, $p < .001$, $d = .66$. For the scenario sets used in Study 1b, a repeated measures ANOVA revealed a significant permissibility difference as well, $F(1, 118) = 15.6$, $p < .001$, partial $\eta^2 = .12$, (mean low = 4.21, $SD = 1.97$; mean high = 4.01, $SD = 2.01$). The main effect of scenario set was significant $F(2, 236) = 7.20$, $p = .001$, partial $\eta^2 = .058$, (mean set 1b = 4.26, $SD = 1.99$; mean set 2b = 3.87, $SD = 1.99$; mean set 3b = 4.20, $SD = 1.98$), as was the interaction of scenario set and condition $F(2, 236) = 4.50$, $p = .012$, partial $\eta^2 = .037$ (mean set 2 low = 4.74, $SD = 2.01$, mean set 2 high = 4.39, $SD = 2.14$; mean set 3 low = 4.38, $SD = 2.01$; mean set 3 high = 4.09, $SD = 1.93$; mean set 4 low = 4.95, $SD = 1.87$, mean set 4 high = 4.34, $SD = 1.93$), suggesting that some scenarios were rated as having bigger permissibility differences than others (see Table 3).

5.3. Discussion

The data support the notion that participants use their perceived likelihoods in forming their moral judgments, even when attention is not explicitly called to those likelihoods. It is possible that questions about likelihood may still have somewhat increased participants' attention to their perceived likelihoods in the previous studies, but this study suggests that perceived likelihoods can influence moral judgment even in the absence of such attentional focusing.

In the first four studies, we focused on the perceived likelihood of the protagonist's action being fatal to the one individual in each scenario. As we indicated earlier, in the moral dilemmas used, we also stipulated that the action causing the death of the one individual would save a larger group of five people. This raises the question of whether participants see the protagonist's action as 100% likely to save the five, for a divergence in the perceived likelihood of this outcome could have implications for participants' moral judgments as well. In the next study, we explored whether participants

Table 3
Rated permissibility between scenarios sets that vary along perceived likelihood of the one dying despite the death being stipulated when not asked probability questions (Study 2)

Scenario Set	Study 1a Set			Study 1b Set		
Version	Less Likely (divert trolley)	More Likely (push man)	Sig. Diff.	Less Likely (e.g., sever foot)	More Likely (e.g., sever neck)	Sig. Diff.
Permissibility	4.29	2.96	***	4.21	4.01	***

Note. ***$p < .001$.

substitute another perceived likelihood for a stipulated outcome that is generally present in trolley problems—the five being saved.

6. Study 3

6.1. Procedure

One hundred and twenty-one participants located within the United States, recruited via Amazon's Mechanical Turk, completed the study (M_{age} = 33.1, SD = 9.59; 60.3% female). Three pairs of scenarios were created to vary along the perceived likelihood that the five would be saved despite explicit stipulation that the five would indeed be saved. In Set 1, a protagonist could drop either a bicycle or a granite block onto a track to stop a trolley from running over five; in Set 2, a protagonist could use either fishing line or climbing rope to rescue a group of five climbers; and in Set 3, a protagonist intent on keeping a bear from mauling five climbers could scare it off by either throwing pebbles at it or shooting it with a tranquilizer dart. (See Supplementary Materials for scenarios.) In both cases, it was explicitly stated that, should the protagonist decide to act, the action would save the five climbers and kill one individual. Study 3 used the same procedure as Study 1b, asking about the likelihood of the one dying as a result of the protagonist's actions, the likelihood of the five being saved, and the permissibility of the action for each scenario.

6.2. Results

We explored whether varying the perceived likelihood of the five being saved by an action would result in permissibility differences between the scenarios. A repeated measures ANOVA found that varying the perceived likelihood of the five being saved by an action resulted in permissibility differences within scenario pairs, $F(1, 120) = 30.8$, $p < .001$, partial $\eta^2 = .20$ (mean low = 3.75, SD = 1.97; mean high = 4.20, SD = 1.92). There was also a main effect of set, such that scenario pairs varied in permissibility from one another, $F(2, 240) = 10.7$, $p < .001$, partial $\eta^2 = .08$, (mean set 1 = 3.22, SD = 1.86; mean set 2 = 3.86, SD = 1.90; mean set 3 = 3.85, SD = 1.95). A significant interaction between scenario set and condition, $F(2, 240) = 4.67$, $p = .010$, partial $\eta^2 = .04$ (mean set 1 low = 4.06, SD = 1.95, mean set 1 high = 4.38, SD = 1.76; mean set 2 low = 3.69, SD = 1.89; mean set 2 high = 4.02, SD = 1.90; mean set 3 low = 3.51, SD = 2.05, mean set 3 high = 4.19, SD = 1.80), suggested that the permissibility difference was much more pronounced in set 3 than in sets 1 or 2.

We next verified that the scenarios had indeed varied the perceived likelihood of the five being saved by the action. Our manipulation did influence the perceived likelihood that the five would be saved, $F(1, 120) = 24.8$, $p < .001$, partial $\eta^2 = .17$, (mean low = 70.7, SD = 30.8; mean high = 79.7, SD = 24.2). There was a main effect of perceived likelihood of the five being saved between sets of scenarios, $F(2, 240) = 19.1$,

$p < .001$, partial $\eta^2 = .14$ (mean set 1 = 80.8, *SD* = 25.8; mean set 2 = 73.7, *SD* = 27.4; mean set 3 = 71.1, *SD* = 29.9) and again there was an interaction between set and condition, $F(2, 240) = 7.73$, $p = .001$, partial $\eta^2 = .06$ (mean set 1 low = 78.6, *SD* = 27.7, mean set 1 high = 82.9, *SD* = 23.8; mean set 2 low = 70.3, *SD* = 29.9; mean set 2 high = 77.2, *SD* = 24.3; mean set 3 low = 63.2, *SD* = 32.8, mean set 3 high = 79.1, *SD* = 24.2). As with the permissibility ratings, the largest effect was observed for the third set.

Next we examined whether perceived likelihoods of the one dying varied within scenario pairs. The repeated measures ANOVA revealed that they did not, $F(1, 120) = .88$, $p = .35$, partial $\eta^2 = .007$ (mean low = 89.0, *SD* = 16.6; mean high = 89.5, *SD* = 16.5). The perceived likelihood of the one dying as a result of the action did, however, vary between scenario sets, $F(2, 240) = 3.86$, $p = .023$, partial $\eta^2 = .03$ (mean set 1 = 91.0, *SD* = 14.9; mean set 2 = 89.1, *SD* = 16.0; mean set 3 = 87.6, *SD* = 18.4). There was no interaction between scenario set and perceived likelihood of the one dying, $F(2, 240) = 1.19$, $p = .31$, partial $\eta^2 = .010$.

Next, a bootstrap mediation analysis with 10,000 resamples revealed that differences in the perceived likelihood of the five dying partially mediated the permissibility differences between pairs of scenarios (mediated effect $ab = .21$, $p < .001$, 95% CI [.12, .36]; direct effect $c' = .24$, $p < .001$, 95% CI [.14, .28]; 47% mediated). These mediation models specified random intercepts and slopes for subjects.

Because we had observed an interaction, each pair of scenarios was analyzed for differences in the two perceived likelihoods and permissibility with paired samples *t* tests. In Set 1, participants did not report dropping the granite block as more likely to save the five (*M* = 82.9%, *SD* = 23.8) than dropping the bicycle (*M* = 78.6%, *SD* = 27.7), $t(120) = 1.77$, $p = .080$, $d = .17$, though there was an observed permissibility difference: Action in the scenario written to be intuitively more likely to save the five was seen as more permissible, $t(120) = 2.89$, $p = .005$, $d = .17$ (mean block = 4.38, *SD* = 1.76; mean bicycle = 4.06, *SD* = 1.95). The perceived likelihood of the one dying marginally differed between the two scenarios, $t(120) = 1.79$, $p = .075$, $d = .12$ (mean block = 90.1%, *SD* = 15.2; mean bicycle = 91.9%, *SD* = 14.7). In both scenarios, participants did not take the death of the one to be 100% likely to occur, nor did they assign a likelihood of 100% to the five being saved.

In Set 2, participants reported that securing the five climbers with climbing rope was significantly more likely to save them (*M* = 77.2%, *SD* = 24.3) than securing them with fishing line (*M* = 70.3%, *SD* = 29.9), $t(120) = 2.94$, $p = .004$, $d = .25$. A difference in permissibility paralleled the difference in the perceived likelihood of the five being saved: action in the scenario perceived as more likely to save the five was seen as more permissible, $t(120) = 3.12$, $p = .002$, $d = .17$ (mean climbing rope = 4.02, *SD* = 1.90; mean fishing line = 3.69, *SD* = 1.89). The perceived likelihood of the one dying did not differ between the two scenarios, $t(120) = .175$, $p = .86$, $d = .05$ (mean climbing rope = 89.0%, *SD* = 15.5; mean fishing line = 89.2%, *SD* = 16.6). Neither scenario produced a judgment of 100% likelihood of the one dying or a judgment of 100% likelihood of the five being saved.

In Set 3, participants reported that shooting the bear with a tranquilizer gun was significantly more likely to save the five climbers (M = 79.1%, SD = 24.2) than throwing pebbles at it (M = 63.2%, SD = 32.8), $t(120) = 5.70$, $p < .001$, $d = .55$. A difference in permissibility paralleled the difference in the perceived likelihood of the five being saved: action in the scenario perceived as more likely to save the five was seen as more permissible, $t(120) = 6.08$, $p < .001$, $d = .35$ (mean tranquilizer = 4.19, SD = 1.80; mean pebbles = 3.51, SD = 2.05). The perceived likelihood of the one dying did not differ between the two scenarios (mean tranquilizer = 87.7%, SD = 18.8; mean pebbles = 87.6%, SD = 18.1), $t(120) = .190$, $p = .850$, $d = .01$. Participants did not take the stipulated outcome of the one dying to be 100% likely, nor did they assign a 100% chance to the stipulated outcome that the five would be saved.

6.3. Discussion

Across the three sets of scenarios, despite the saving of the five being explicitly stipulated, participants reported not only that the likelihood of the outcome was well below 100%, but also that there was a difference in likelihood between pairs of scenarios. Furthermore, an increased perceived likelihood of the five being saved corresponded to an increased degree of permissibility of the action (just as an increased perceived likelihood of the one dying corresponded to a decreased degree of permissibility in Study 1b).

The next study examined whether, as with the pair of studies exploring the perceived likelihood of the one individual's death, the differences in permissibility would persist without focus being directed to the likelihoods. That is, in the absence of questions regarding likelihoods, would participants nevertheless find action in the scenario versions previously found to be more likely to result in the deaths of the five to be less permissible? Because Set 1 was not rated as having a significant difference in perceived likelihood of the five being saved, and Set 2 had modest effect size, relative to both Set 3 and the scenarios sets in Study 1b, Set 3 may be the most informative test case going forward.

7. Study 4

7.1. Procedure

One hundred and nineteen participants located in the United States were recruited via Mechanical Turk (M_{age} = 31.4, SD = 8.59; 55.4% female). Study 4 repeated the procedure from Study 3, although with questions relating to likelihood removed, leaving just the single permissibility question for each scenario.

7.2. Results

Permissibility differences between scenario sets that vary in perceived likelihood, but do not probe likelihood, were analyzed, using a repeated measures ANOVA. The pairs of

scenarios, written to vary in perceived likelihood of the five being saved only, did not show an overall difference when not probing for likelihood, $F(1, 118) = .74$, $p = .39$, partial $\eta^2 = .006$ (mean low = 3.64, $SD = 1.75$; mean high = 3.67, $SD = 1.77$). The main effect of scenario set was significant $F(2, 236) = 5.60$, $p = .005$, partial $\eta^2 = .045$, (mean set 1 = 3.80, $SD = 1.79$; mean Set 2 = 3.60, $SD = 1.75$; mean set 3 = 3.57, $SD = 1.74$), and there was a significant interaction of scenario set and condition $F(2, 236) = 3.01$, $p = .05$, partial $\eta^2 = .025$. This suggested that some scenarios were rated as having bigger permissibility differences than others.

Because we again found an interaction, we examined the sets individually. Permissibility differences were observed in one of three scenario sets. In Set 3, where the protagonist could save climbers from a bear by using either a tranquilizer gun or pebbles, the difference in permissibility remained significant, $t(118) = 2.36$, $p = .02$, $d = .10$ (mean tranquilizer = 3.66, $SD = 1.72$; mean pebbles = 3.48, $SD = 1.76$). Permissibility did not, however, differ in Set 1, where either a bicycle ($M = 3.83$, $SD = 1.79$) or granite block ($M = 3.76$, $SD = 1.80$) would be dropped, $t(118) = .97$, $p = .33$, $d = .04$, or in Set 2, where climbers would either be secured using climbing rope ($M = 3.59$, $SD = 1.78$) or fishing line ($M = 3.61$, $SD = 1.72$), $t(118) = .35$, $p = .72$, $d = .01$.

7.3. Discussion

The findings show that, even when participants were not explicitly asked about likelihood, the perceived likelihood of an action's saving five can override stipulated outcomes and influence moral judgments. These results suggest that having participants think specifically about the likelihood of the stipulated outcome of action did sensitize them to those differences, and likely enhanced the rated permissibility differences in the prior study. Nonetheless, a difference can emerge even without attention being explicitly directed to the likelihoods.

What explains the fact that we found significant differences with respect to moral permissibility judgments in Set 3, but not Set 1 and Set 2? It is notable that when participants were presented with Set 1 in Study 3, we found an insignificant difference in participants' perceived likelihoods of outcomes. Perhaps surprisingly, participants there rated the likelihood of the five being saved to be very similar whether a bicycle or a 100-ton granite block was released onto the track. Participants did offer significantly different likelihood judgments in Set 2, but notably, these differences were not as great as those in Set 3. It is possible that in cases like Set 1, where participants are not inclined to offer very different likelihood judgments even when asked about likelihood of outcomes, simply being asked can enhance the salience of even minimal perceived likelihood differences in answering questions about moral permissibility. It is also possible that when asked about the likelihood of saving the five, other perceived likelihoods are brought to salience. (We explore the possibility in Study 6 that participants are responding to a number of different perceived probabilities related to each scenario.) In any case, it is perhaps not surprising that the set which revealed the greatest difference in perceived likelihoods in Study 3, namely Set 3, generated the greatest difference in moral judgments when participants were not asked about likelihood at all.

Until now, we have used continuous ratings of permissibility. Participants use such a scale without complaint, and the responses do appear to covary with their perception of various morally relevant likelihoods. However, the meaning of such a continuum is not entirely clear. What, for example, does it mean for an action to be rated halfway between permissible and impermissible? In many contexts, people treat the words "permissible" and "impermissible" in a way parallel to the words "legal" and "illegal," and yet it is not at all clear what it would mean to be halfway between "legal" and "illegal." Moreover, most conceptual analyses of permissibility and impermissibility treat these concepts as binary. For example, "impermissible" is most often treated as synonymous with "forbidden" or "something one ought not to do," which express non-scalar notions. And the very few defenses of scalar conceptions of rightness or permissibility, such as Lockhart (2000) and Peterson (2013), recognize that the scalar view is non-standard. Thus, there is at least an important concept of (im)permissibility that is non-scalar.

Now it does not follow from the fact that there is a non-scalar concept of (im)permissibility that people are unable to use the words "permissible" and "impermissible" to express scalar notions. We believe that this is what is going on when participants are presented with Likert permissibility scales. For example, they may mean by "less permissible" something like "involving a more serious moral infraction" or "its being *worse* (or causing *more harm*, or infringing a *more stringent* right) to do one thing rather than another" (such as to kill rather than steal a piece of gum). At the same time, the standard conception of moral permissibility in moral theory corresponds to the standard (and, so far as we are aware, the only) treatment of legality in legal theory. All felonies are illegal, and none is more (or less) illegal than any other. But clearly some felonies are worse than others, cause more harm than others, or infringe more stringent rights than others.

If there is an important concept captured by "permissible" that is binary, and judgments of permissibility are affected by participants' substituted perceived likelihoods of outcomes, then we should expect to see this reflected when using a binary measure of permissibility. For the next study we adopted a binary (yes/no) permissibility evaluation and explored whether differences in the perceived likelihoods in scenarios can influence binary permissibility judgments. For this study, we also tested whether the permissibility judgment differences would be apparent when each participant rated only one of the pair of scenarios, preventing their making any comparative likelihood or permissibility rating based on the small alterations within each pair. We also did not ask for explicit likelihood judgments, to avoid the possibility of priming driving any difference in permissibility judgments.

8. Study 5

8.1. Procedure

One hundred and twenty-two participants located in the United States were recruited via Mechanical Turk (M_{age} = 35.5, SD = 11.7; 63.9% female). Study 5 used Set 3 from Studies 3/4 to investigate binary permissibility judgments. Each participant saw only one

member of the pair of scenarios. Rather than respond on a 7-point scale, participants made a binary yes/no judgment regarding the permissibility of the protagonist's action.

8.2. Results

Of the 61 participants who saw the version of the scenario in which the protagonist can save the five climbers from the bear by shooting it with a tranquilizer gun, 32 (52.5%) thought it would be permissible to perform the action. Of the 61 participants who saw the version in which the protagonist can save the five climbers from the bear by throwing pebbles at it, only 21 (34.4%) thought it would be permissible to act. The difference in the distributions of judgments was significant, $\chi^2(N = 122) = 4.06$, $p = .044$, OR = 2.10, 95% CI [1.01, 4.36], suggesting that participants were more likely to think it permissible to kill the one to save the five when the bear would be shot with the tranquilizer gun than when pebbles would be thrown at the bear, despite both versions explicitly stating that the action would save the five climbers.

8.3. Discussion

In this study, perceived likelihoods were found to override explicitly stated outcomes and to affect binary permissibility judgments, just as in previous studies they had affected scalar permissibility judgments. It was also apparent that the effect emerges in a between-subjects design just as it had in within-subjects designs. And, as before, the effect emerges even when there is no explicit question about the likelihood of outcomes.

9. Study 6

So far we have examined the way that people bring their own perceived likelihoods to bear on moral decisions with respect to pairs of written scenarios where the chances of the protagonist's action leading to the death of one individual, or to the saving of five individuals, are plausibly different. However, this does not exhaust (potentially morally relevant) likelihoods. In particular, one can distinguish between the likelihoods of outcomes consequent upon inaction as well as upon action. The likelihoods in the previous studies reflect the likelihood of certain events occurring should a particular action be performed. What if one varied the likelihood of events occurring should the action *not* be performed? Would participants judge it to be less permissible to kill one to save five if the five had some chance of surviving anyway? Or would they judge it to be more permissible to act if the single person was already at some risk of death? This study explores these questions.

9.1. Procedure

One hundred and twenty-seven participants located within the United States were recruited via Mechanical Turk (M_{age} = 35.0, SD = 10.4; 63.0% female). One set from

Studies 1/2 was used, where our intent was to manipulate perceived likelihoods of the one dying, as was one set from Studies 3/4, where the intent was to manipulate perceived likelihoods of the five being saved. Two new scenario pairs were created to vary along perceived likelihoods relating to inaction. The first set stipulated that the one would survive if the action was not taken, but varied along the perceived likelihood that this would occur. The second was intended to let intuitions vary about the likelihood that the five would die if the action was not taken. Thus, in this study, we inquired about the perceived likelihoods of all four of these outcomes: (a) how likely the one was to live if the protagonist did not act, (b) how likely the one was to die if the protagonist acted, (c) how likely the five were to die if the protagonist did not act, and (d) how likely the five were to live if the protagonist acted. We also asked about permissibility. This allowed us to explore whether all of these perceived likelihoods have ramifications for moral judgments, as well as the extent to which they can be altered independently, even in scenarios where the outcomes are stipulated.

We used Set 1 from Studies 1b and 2, in which the action would either kill the one by severing his neck or his foot, which varied the perceived likelihood that the one would die as a result of the action. The next set varied the perceived likelihood that the one would die if the action were not taken. In one version, an individual miner has his foot stuck in rubble, and in the other he is trapped and severely injured. Although we asserted that the one would live in both cases, it seems intuitively that the one whose foot is stuck is more likely to do so. As a result, it might seem that blasting the hole in the rubble he is trapped in to rescue five other miners is more permissible in the latter case.

We used Set 3 from Studies 3 and 4, in which the five could be saved from a bear by throwing pebbles at the bear or shooting it with a tranquilizer dart, thus varying the perceived likelihood that the five would be saved with action. The final set varied the perceived likelihood that the five would survive despite inaction: Although in both members of the pair the death of the climbers is stipulated, it might be thought less likely that the climbers will die when they fall 10 feet onto a flat rock surface, compared to when they plummet 1,000 feet onto jagged rocks. This may make killing one person to save five seem less permissible.

Participants saw each version of each scenario, randomly ordered. After reading each scenario, they answered four questions regarding likelihood and one regarding permissibility. Two of the likelihood questions concerned the likelihood of the one dying and the five being saved if the protagonist decides to act—these were the same questions we asked in Studies 1 and 3. Two more likelihood questions, relating to inaction, were also asked: one about the likelihood of death of the one individual if the protagonist decides not to act (e.g., *If Jason decides not to rapidly traverse the narrow ledge, how likely is it that the man on the ledge will* die? 0–100%), the other about the likelihood of the five surviving if the protagonist decides not to act (e.g., *If Jason decides not to rapidly traverse the narrow ledge, how likely is it that five will survive*? 0–100%).

9.2. Results

9.2.1. Set 1

This set was intended to vary the perceived likelihood of the one dying if the protagonist acts. Indeed, replicating the finding from Study 1b, severing the neck was rated more likely to lead to death (M = 92.5%, SD = 19.4) than severing the foot (M = 77.8%, SD = 27.3), $t(126) = 5.61$, $p < .001$, $d = .62$ (see Table 4). Fifty out of the 127 participants thought severing the foot had a greater than 95% chance of actually killing the one, whereas 99 out of the 127 thought severing the head did. A difference in permissibility paralleled this difference in perceived likelihood: Action in the version of the scenario rated more likely to lead to the one individual's death was rated less permissible (mean neck = 4.63, SD = 2.02; mean foot = 4.94, SD = 1.91), $t(126) = 2.61$, $p = .010$, $d = .16$.

Other perceived likelihoods did not differ between the two scenarios of Set 1. As in Study 1b, the perceived likelihood of the five being saved if action was taken did not differ, $t(126) = 1.28$, $p = .20$, $d = .11$ (mean neck = 89.9%, SD = 21.2; mean foot = 92.0%, SD = 17.1). The perceived likelihood of the one dying if action was not taken also did not differ, $t(126) = .872$, $p = .39$, $d = .08$ (mean neck = 19.3%, SD = 30.1; mean foot = 17.0%, SD = 27.4). The perceived likelihood of the five surviving if action was not taken did not differ either, $t(126) = 1.24$, $p = .22$, $d = .11$ (mean neck = 21.1%, SD = 30.9, mean foot = 17.9%, SD = 27.3). Outcomes that were stipulated to occur were assessed as less than 100% likely, and outcomes that the scenario suggested would not occur (e.g., the single person dying on the tracks even if no action is taken) were rated as having some non-zero chance of occurring.

9.2.2. Set 2

This set was intended to vary the perceived likelihood of the five being saved if the protagonist acts. As in Study 4, shooting the bear with a tranquilizer gun was again thought to be more likely to save the five (M = 83.2%, SD = 21.0) than throwing pebbles at it (M = 72.1%, SD = 28.6), $t(126) = 5.14$, $p < .001$, $d = .44$. Forty-one participants thought it more than 95% likely that throwing pebbles at the bear would prevent it from attacking, while 49 participants thought the same about shooting the bear with the tranquilizer gun. A difference in permissibility paralleled this difference in perceived likelihood: The action judged more likely to save the five was deemed more permissible, $t(126) = 2.36$, $p = .020$, $d = .12$ (mean tranquilizer = 4.06, SD = 1.97; mean pebbles = 3.83, SD = 1.94).

Other perceived likelihoods did not differ between the scenarios of Set 2. Again, the perceived likelihood of the one dying if action was taken did not differ, $t(126) = .661$, $p = .51$, $d = .04$ (mean tranquilizer = 89.3%, SD = 18.6; mean pebbles = 88.5%, SD = 20.8). The perceived likelihood of the one dying if action was not taken did not differ either, $t(126) = .350$, $p = .73$, $d = .02$ (mean tranquilizer = 11.8%, SD = 21.3; mean pebbles = 12.3%, SD = 22.6), nor did the perceived likelihood of the five surviving if action was not taken, $t(126) = .718$, $p = .47$, $d = .07$ (mean tranquilizer = 28.1%, SD = 29.2; mean pebbles = 26.1%, SD = 28.0). Outcomes that were stipulated to occur

Table 4
Rated likelihoods and permissibility between scenario sets that vary along four intuitive probabilities, despite the outcome being stipulated for each (Study 7)

Scenario Set	Set 1			Set 2			Set 3			Set 4		
Version	Less Likely (sever foot)	More Likely (sever neck)	Sig. Diff.	Less Likely (pebb-les)	More Likely (tranq-uilizer gun)	Sig. Diff.	Less Likely (foot trapped)	More Likely (body trapped)	Sig. Diff.	Less Likely (fall 10 feet)	More Likely (fall 1000 feet)	Sig. Diff.
Rated likelihood of the one dying if action taken (%)	**77.8**	**92.5**	***	88.5	89.3	***	88.9	90.3	ns	80.1	89.3	ns
Rated likelihood of the five being saved if action taken (%)	92.0	89.9	ns	**72.1**	**83.2**	***	84.6	82.9	ns	84.1	80.6	†
Rated likelihood of the one dying w/out action (%)	17.0	19.3	ns	12.3	11.8	ns	**29.1**	**39.6**	***	13.8	14.7	ns
Rated likelihood of the five surviving w/out action (%)	17.9	21.1	ns	26.1	28.1	ns	20.7	20.1	ns	**37.8**	**19.4**	***
Permissibility	4.94	4.63	*	3.83	4.06	*	4.42	4.89	***	3.64	4.31	***

Note. Bolded likelihoods reflect the probability intended to intuitively vary in each scenario. †$p < .1$, *$p < .05$, **$p < .01$, ***$p < .001$.

were assessed as less than 100% likely, and outcomes that were stipulated as not occurring were assessed as more than 0% likely.

9.2.3. Set 3

This set was intended to vary the perceived likelihood of the death of the one, if the protagonist does not act to save the five. In the mining scenario, participants thought the one was more likely to die when trapped and severely injured (M = 39.6%, SD = 30.4) than when only his foot is stuck (M = 29.1%, SD = 33.9), $t(126) = 3.40$, $p < .001$, $d = .33$. A difference in permissibility paralleled this difference in perceived likelihood: Action leading to the one's death in the scenario in which the one was perceived as less likely to survive in the absence of action was judged more permissible, $t(126) = 4.36$, $p < .001$, $d = .24$ (mean severe = 4.89, SD = 1.91, mean foot = 4.42, SD = 1.97).

Other perceived likelihoods did not differ between the pairs of Set 3. The perceived likelihood of the one dying if action was taken did not differ, $t(126) = .587$, $p = .56$, $d = .07$ (mean severe = 90.3%, SD = 20.1; mean foot = 88.9%, SD = 22.6). The perceived likelihood of the five being saved if action was taken did not differ, $t(126) = .873$, $p = .38$, $d = .08$ (mean severe = 82.9%, SD = 23.5; mean foot = 84.6%, SD = 20.8), nor did the perceived likelihood of the five surviving if action was not taken, $t(126) = .280$, $p = .78$, $d = .02$ (mean severe = 20.1%, SD = 26.3; mean foot = 20.7%, SD = 27.7). Stipulated outcomes were not judged as either definitely occurring or definitely not occurring.

9.2.4. Set 4

The final set was intended to vary the perceived likelihood of the five surviving without the protagonist acting to secure their bridge. Despite the stipulation that falling would lead to the climbers' deaths in either scenario, participants reported thinking that if the climbers fell 10 feet onto a flat rock surface they were more likely to survive (M = 37.8%, SD = 36.2), compared to the scenario where they would plummet 1,000 feet onto jagged rocks (M = 19.4%, SD = 27.6), $t(126) = 5.16$, $p < .001$, $d = .57$. Judgments of permissibility significantly differed, inversely to the difference in perceived likelihood of the five surviving, $t(126) = 5.20$, $p < .001$, $d = .34$ (mean short fall = 3.64, SD = 2.01; mean long fall = 4.31, SD = 1.93).

Perceived likelihoods other than the likelihood of the five surviving even without action did not differ between scenarios. Participants did not think that the pair of scenarios varied along likelihood of the one dying if action was taken, $t(126) = .594$, $p = .55$, $d = .05$ (mean short fall = 90.1%, SD = 16.2; mean long fall = 89.3%, SD = 17.7). Nor did they take the scenarios to differ in likelihood of the one dying if action was not taken, $t(126) = .395$, $p = .69$, $d = .04$ (mean short fall = 13.8%, SD = 25.3; mean long fall = 14.7%, SD = 25.1). However, there was a marginal difference in the perceived likelihood of the five being saved, $t(126) = 1.90$, $p = .060$, $d = .16$ (mean short fall = 84.1%, SD = 19.9; mean long fall = 80.6%, SD = 23.7), suggesting that the perceived likelihood of the five surviving due to inaction contributed to the perceived likelihood of the five being saved. Outcomes that were stipulated to occur were assessed as

less than 100% likely, and those that participants might be expected to understand as not occurring were judged as more than 0% likely.

In this study, where we find differences in the perceived likelihoods of the one dying and the five surviving without action, and the one surviving and the five dying with action, and associated differences in the permissibility of action, we can explore the relationship between perceived likelihoods and permissibility. That is, we can test whether the alterations in perceived likelihood are mediating the differences in permissibility. Bootstrap mediational analysis (Tingley et al., 2014) was used to investigate this for each of the four sets, with random intercepts entered for each subject.

There was a significant indirect effect of Set 1 (which was designed to manipulate perceived likelihood of the action actually killing the single person) on permissibility through probability, $ab = .15$, $p < .001$, 95% CI [.04, .28] (10,000 resamples). The direct effect was no longer significant, $c' = .16$, $p = .21$, 95% CI [−.09, .40], suggesting that the difference in perceived likelihood of the one actually dying fully mediated permissibility differences. While the effect sizes reported throughout these studies are small, their magnitude may be misleading due to the large variance in perceived likelihoods—despite change in perceived likelihoods being small relative to this range, their movement entirely mediates the permissibility difference observed here.

For Set 2 (which manipulated perceived likelihood that the action would indeed save the five), there was a significant indirect effect on permissibility through probability, $ab = .21$, $p < .001$, 95% CI [.12, .33] (10,000 resamples). The direct effect was no longer significant, $c' = .01$, $p = .89$, 95% CI [−.18, .20], suggesting that the difference in perceived likelihood of the five actually being saved by the action fully mediated permissibility differences between the two scenarios.

There was a significant indirect effect of Set 3 (which manipulated the perceived likelihood that the single person might die even without action) on permissibility through probability, $ab = .12$, $p < .001$, 95% CI [.04, .22] (10,000 resamples). The direct effect remained significant, $c' = .35$, $p < .001$, 95% CI [.14, .56], suggesting that the difference in perceived likelihood of the one dying regardless of whether action was taken partially mediated permissibility differences. This may suggest that some counterfactual probabilities are hard to explicitly convey, despite their effect on permissibility.

Finally, there was a significant indirect effect of Set 4 (which manipulated the perceived likelihood of the five surviving even without action) on permissibility through probability, $ab = .29$, $p < .001$, 95% CI [.16, .45] (10,000 resamples). The direct effect remained significant, $c' = .39$, $p < .001$, 95% CI [.14, .63], suggesting that the difference in perceived likelihood of the five surviving regardless of whether action was taken partially mediated permissibility differences. This provides further evidence that the rated likelihood may not necessarily perfectly capture the intuitive likelihood's impact on permissibility judgments.

9.3. Discussion

Study 6 found that each scenario we had written to specifically vary one of four perceived likelihoods was associated with corresponding differences in participants'

judgments about likelihood as well as permissibility. This result supports the hypothesis that intuitions regarding a variety of kinds of likelihoods can be substituted for scenario stipulations. Stipulated outcomes were never taken to have a 100% probability of occurring. The action that was described as saving five people was not seen as perfectly likely to have that effect, nor was the action described as omitting to save them seen as leaving them to their certain death. Similarly, the action that was described as resulting in the death of the single individual was not seen as 100% likely to have that effect, nor was the corresponding omission seen as any kind of guarantee that the individual would survive. Moreover, the fact that differences in each type of perceived likelihood were associated with differences in permissibility judgments supports the hypothesis that intuitions about each type of likelihood can have an effect on moral judgment. In all cases, we found that the perceived likelihoods did mediate, either partially or fully, the impact on permissibility. This reinforces the notion that intuitions about actual likelihoods, even when the relevant outcomes are stipulated to occur, can alter moral judgments.

Our studies so far suggest that perceived likelihoods are at least partial drivers of permissibility judgments. We can now revisit the hypothesis we entertained in our discussion of Study 1a, viz. that perceived intention may influence probability judgments. Our final study explored whether such probability differences are relevant not only in consideration of the details of scenarios, but even occur as a result of the underlying moral principles the scenarios are thought to convey. The question is whether, regardless of the details, describing harm as intended can make it seem more likely to occur than describing it as merely foreseen.

10. Study 7

10.1. Procedure

One hundred and seventeen adults located in the United States were recruited via Mechanical Turk (M_{age} = 36.7, SD = 12.0; 60.7% female). Study 7 presented participants with two scenarios, written to differ along the foreseen/intended distinction abstractly, to avoid any potential effect of differing details on perceived likelihoods. Participants read both scenarios, and then two questions.

Foreseen:

Plan X

Imagine you are told that Jason has a plan to save a group of children from a terrorist attack, and that a foreseen, but not intended, consequence of Jason carrying his plan out is that Thomas, a bystander, is killed.

Intended:

Plan Y

Imagine you are told that Harry has a plan to save a group of children from a terrorist attack, and that Harry intends to kill a bystander, Robert, because for Harry's plan to work it is necessary that Robert is killed.

Participants responded to "*Which plan is more likely to result in the death of the bystander?; −7 = Plan X is more likely than Plan Y to result in the bystander's death to 7 = Plan Y is more likely than Plan X to result in the bystander's death*" and "*Which of the two plans is morally worse?; −7 = Plan X is much worse than Plan Y to 7 = Plan Y is much worse than Plan X*."

10.2. Results

Responses to the scenario set were analyzed for differences in relative perceived likelihoods and relative moral acceptability with one–sample t tests comparing responses to 0 (equal). Participants reported thinking that the bystander was significantly more likely to die in the intended plan than in the foreseen plan $t(116) = 6.54$, $p < .001$, $d = .60$ ($M = 1.93$; $SD = 3.20$). A difference in relative moral acceptability paralleled this difference in perceived likelihood: The plan describing the intended death of a bystander was judged to be significantly less morally acceptable than the plan describing a foreseen death, where the bystander's death had been perceived to be less likely, $t(116) = 8.92$, $p < .001$, $d = .82$ ($M = 2.99$; $SD = 3.63$). Relative perceived likelihood ratings correlated with relative moral acceptability ratings, $F(1, 115) = 23.4$, $p < .001$, $r = .41$.

10.3. Discussion

We began our investigation in Study 1a with a scenario pair modeled on the classic trolley dilemma. Responses to this dilemma and others like it are sometimes taken to support the claim that something like the Doctrine of Double Effect is tacitly operative in moral judgment. Our final study shows that the foreseen/intended distinction, presented in the abstract and stripped of additional likelihood-relevant details, still conveys likelihood-relevant information that appears to contribute to differences in moral judgments. This is an important result. Together with the results from Studies 1–6, it suggests that judgment patterns consistent with the Doctrine of Double Effect cannot be readily interpreted as showing that something like the DDE is tacitly operative in moral thought, since in many cases such a principle is confounded with differences in perceived likelihoods. Since the details of the scenarios—both those related to and those unrelated to intention—convey differential likelihood information relevant to and potentially impacting moral judgment, moral dilemma tasks designed to test for the DDE face the serious challenge of isolating a pure comparison on the dimension of foreseeability and intention, and not implicitly conveying differential likelihoods of outcomes. This is not to say that the challenge cannot be met, or that something like the DDE is in fact tacitly operative. But more work needs to be done to demonstrate a clear test of the operation of the DDE.

11. Meta-analytic results

A meta-analytic approach was taken to combine results across the studies that were designed to differ only in perceived likelihoods, not in morally relevant factors (namely, Studies 1b and 2–6). We calculated average effect sizes for permissibility differences between pairs of scenarios across studies when participants had been asked to rate likelihoods of the outcomes, and also when they had not been so asked, controlling for the sample size of each study. Studies that asked about likelihoods alongside permissibility had a standardized effect size of .24, $p = .001$ (Study 1b, Study 3, and Study 6). A parallel examination of studies that did not ask about likelihood showed a standardized effect size of .17, $p = .022$ across Study 2, Study 4, and Study 5. The last study required using Chinn's (2000) approach to converting the odds ratio to an effect size.

12. General discussion

In seven studies, we explored the extent to which people accept stipulated outcomes in hypothetical moral scenarios, how this can differ between scenarios, and how such differences in intuitive likelihoods can result in different judgments of permissibility. Study 1a found that a scenario set representative of those used to study the Doctrine of Double Effect differed not only in terms of whether the harm was intended or foreseen, but also in the intuitive likelihood of death of the one individual, as well as in the intuitive likelihood of the five being saved, despite both of these outcomes being stipulated. Study 1b sought to test whether scenario pairs which held everything else constant and differed only with respect to the intuitive probability that the individual would die if the protagonist acted would produce differences in probability and permissibility ratings, and found, across multiple scenarios, that they do. Studies 1c and 1d supported the idea that participants reporting less-than-certain likelihoods was not an artifact of our experimental method.

Study 2 verified that differences in permissibility ratings could arise in the absence of questions regarding the intuitive likelihood of the death of the one individual, though these questions may have sensitized participants to intuitive differences in likelihood. Study 3 examined scenarios written to isolate the difference in intuitive likelihood of the five being saved, rather than of the one dying, and showed that this difference alone could also result in a difference in permissibility ratings between scenarios. Study 4 demonstrated that even in the absence of questions regarding the intuitive likelihood of the five being saved, differences along this dimension could result in divergent permissibility ratings. Study 5 replicated this result while presenting each participant with only one member of a pair of scenarios and with a binary yes/no permissibility judgment, suggesting that our findings hold even when participants are not exposed to both versions of a scenario, without any sensitization due to asking about intuitive likelihoods, and with a response measure that better reflects most philosophers' view of the non-scalarity of

permissibility. Study 6 explored how two other probabilities present in many moral dilemmas—the likelihood of the one dying and of the five surviving without the protagonist's intervention—could likewise affect permissibility ratings. Study 7 extended our findings about the connection between intuitive likelihood and moral judgment, suggesting that the foreseen/intended distinction similarly implicates differential judgments of likelihood, and that this can have a corresponding effect on moral judgment.

These findings have implications that are both methodological and conceptual. It seems clear that people do not understand the scenarios in precisely the way they are intended. We have shown that interventions that we assert will save five people may not be seen as necessarily doing so. Nor will the steps taken to, perhaps, save those five necessarily harm the single individual. In fact, the gap between how scenarios are conceived and how they are understood may go beyond this. When we assert that five people are helpless on one of the tracks the trolley might travel, perhaps we are wrong about there being precisely five of them. And perhaps those five (or so) people, ones we have imagined to be just random innocents, are actually seen as complicit in their own fate—who, after all, hangs out on tracks down which trolleys can readily travel? Such a view is actually familiar to those who have posed moral scenarios to people only to have them quibble with the premises. "Why not yell ahead and alert the people on the tracks?" or "Perhaps the conductor should just put the switch in the middle position and safely derail the trolley." Such views are not irrational, though they do pose challenges to researchers. People are, generally, wise to take their prior probabilities into account in decision making, and, in other contexts, are often criticized for doing so insufficiently (Tversky & Kahneman, 1974).

When we ask participants whether it is worth sacrificing one to save five in detailed scenarios, it appears that many understand the questions in terms of relative risk. Differences between individuals in moral judgments may be due in some part to the extent to which they accept the assertions within the scenario rather than substitute their own intuitive probabilities, how they form the estimate of those probabilities, and then, to some extent, what intuitive moral theory they hold and apply. That last factor—the one of most direct interest to those investigating people's moral reasoning—is more difficult to isolate than one might hope. Simply asserting that everything else is fixed appears insufficient.

As an alternative explanation of our data, it might be suggested not that the intuitive probabilities that participants substitute for stipulations affect their moral judgments, but rather that participants' moral judgments affect the intuitive probabilities that they substitute for stipulations (see Liu & Ditto, 2013). Although some of our studies do not rule out this possibility, others speak against it. Consider, for example, Set 1 in Study 2, which consists of a pair of cases that differ only with respect to which body part of the one individual, neck or foot, will be severed by a trolley that has been diverted in his direction. Although participants are told in both cases that the one individual will die, they judge the likelihood of his dying as falling significantly short of 100% when what is severed is his foot (M = 79.8%), but as falling nearer to 100% when what is severed is his neck (M = 94.5%). If the difference in participants' judgments of degree of moral permissibility of diverting the trolley were driving the difference in their judgments of

the likelihood of the one's death, then something other than the latter difference would need to be driving the former. But what could that be? Apart from the foot/neck difference, the two cases are exactly the same. Were there some *further* factual difference between the cases that might account for the difference in moral judgment (e.g., that in the one case the five are all children while in the other they are all adults, or that in the one case the protagonist made a promise to the five while in the other the protagonist did not), it might be possible to explain the difference in participants' moral judgments in a way that coheres with the assumption that this difference is itself driving the difference in participants' intuitive probabilities. But there is no such *further* factual difference in Set 1 of Study 2. We are, therefore, left with what seems in any event to be the most plausible explanation of participants' moral judgments in these cases: Participants find it easier to justify diverting the trolley at least in part *because* they believe that it is less likely that the one who will experience the severing of a body part will actually die.

Our studies have important implications for the fields of moral psychology and experimental philosophy. Until we can ensure that participants are not substituting their own intuitive probability estimates for experimenters' stipulations, we are not permitted to draw any unqualified conclusions from studies of moral judgments that are based on participants' evaluations of protagonist behavior in vignettes that are simply presented to them on paper or online. Our studies also suggest that we need to be very careful before drawing any conclusions about folk moral intuitions on the basis of moral conversations with people who may well refuse to accept stipulations about outcomes in hypothetical cases. As previously noted, some people (e.g., in classrooms) push back against the stipulations, and we often exclude their reactions from the data for this reason; but our results establish that many participants who do not *explicitly* resist the stipulations in fact resist them anyway (consciously or not). Future studies need to correct for the fact that many participants judge whether a protagonist's behavior is morally better or worse (more or less) morally permissible or impermissible, on the basis of their own sense of how likely it is that certain events will occur or will not occur depending on whether the protagonist acts or does not act. Future work needs to address the best way of making this correction: One solution might be to write scenarios in which intuitive probabilities are matched, and another might be to find a way to present scenarios so that participants accept all the embedded stipulations.

If, as we suspect, participants routinely refuse to accept stipulations that conflict with their antecedent probability expectations, then some doubt might be cast on a wide variety of past vignette-driven experiments, designed to elicit not merely *moral* judgments but other philosophical judgments (e.g., about the compatibility of free will with determinism), causal judgments, linguistic judgments, mental state judgments, and more. The proper design of studies in vignette-driven psychology and experimental philosophy needs to be rethought.

There are at least two more questions raised by our studies that deserve further consideration. The first question is whether further intuitive probabilities not discussed in our studies play some role in participants' moral judgments about the actions of protagonists

in hypothetical scenarios. Participants, for example, might think that there is a non-negligible likelihood of a runaway trolley posing a danger to more people than the six who are usually mentioned in a typical trolley case. They could also include their intuitive likelihood that the various protagonists are on the track in the path of a trolley through their own negligence. It might even be that participants' moral judgments are affected by their estimate of the ex ante likelihood of the hypothetical scenario itself. Shtulman and Tong (2013), for example, found that, as an individual difference variable, regarding extraordinary events as possible (i.e., having a non-zero likelihood) was associated with finding different actions as possibly being permissible. It would certainly be valuable to know just how many, and how extensively, antecedent probability estimates of different types affect moral judgment.

The second question concerns participants' judgments about the morality of risk. To discover which features of the world participants take to be morally significant (and how morally significant they take them to be), researchers ask them questions based on vignettes in which it is stipulated that such-and-such an action or omission *will* (in no uncertain terms) lead to so-and-so outcome (e.g., "If Sam throws a switch, the trolley *will* be diverted"). What, then, do participants really believe about what morality forbids or permits in the presence of greater or lesser risk? If, for example, participants are faced with a scenario in which the protagonist must, in order to save five, choose between imposing a near-100% risk of death on one or imposing a 50% risk of death on two (or, perhaps, a scenario in which the protagonist must choose between a course of action that has a near-100% chance of saving one and a course of action that has a 50% chance of saving two), what will they recommend? Scenarios typically stipulate fixed outcomes, but in the real world there is always uncertainty about the results of acting, or failing to act. An understanding of people's moral reasoning should reflect this.

Acknowledgments

We thank UC San Diego's Frontiers of Innovation Scholars Program and the Dean of Arts and Humanities, Cristina Della Coletta, for supporting our research.

References

Bennis, W. M., Medin, D. L., & Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspectives on Psychological Science*, *5*(2), 187–202.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131.

Christensen, J., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience and Behavioral Reviews*, *36*(4), 1249–1264.

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

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

Greene, J., Cushman, F., Stewart, L., Lowenberg, K., Nystrom, L., & Cohen, J. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language*, *22*(1), 1–21.

Kortenkamp, K. V., & Moore, C. F. (2014). Ethics under uncertainty: the morality and appropriateness of utilitarianism when outcomes are uncertain. *American Journal of Psychology*, *127*(3), 367–382.

Liu, B. S., & Ditto, P. H. (2013). What dilemma? Moral evaluation shapes factual belief. *Social Psychological and Personality Science*, *4*(3), 316–323.

Lockhart, T. (2000). *Moral uncertainty and its consequences*. New York, NY: Oxford University Press.

McIntyre, A. (2004). Doctrine of double effect, The Stanford encyclopedia of philosophy (Winter 2014 Edition), Edward N. Zalta (ed.). Available at http://plato.stanford.edu/archives/win2014/entries/double-effect/

Mikhail, J. (2000). Rawls' linguistic analogy: A study of the "Generative Grammar" model of moral theory described by John Rawls in "A Theory of Justice." PhD dissertation, Cornell University.

Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill?: Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549–557.

Nichols, S., & Knobe, J. (2007). Moral responsibility and determinism: The cognitive science of folk intuitions. *Nous*, *41*(4), 663–685.

Peterson, M. (2013). *The dimensions of consequentialism: Ethics, equality and risk*. Cambridge, MA: Cambridge University Press.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Shtulman, A., & Tong, T. (2013). Cognitive parallels between moral judgment and modal judgment. *Psychonomic Bulletin & Review*, *20*, 1327–1335.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*(5).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124.–1131.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Scenario Texts.