Cognition 209 (2021) 104548

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Sensitivity to shifts in probability of harm and benefit in moral dilemmas

Arseny A. Ryazanov [a,*], Shawn Tinghao Wang [b], Samuel C. Rickless [b], Craig R.M. McKenzie [c], Dana Kay Nelkin [b]

[a] Department of Psychology, University of California, San Diego, United States
[b] Department of Philosophy, University of California, San Diego, United States
[c] Rady School of Management, University of California, San Diego, United States

ARTICLE INFO

Keywords:
Moral cognition
Ethics
Risk
Decision-making
Moral dilemma
Probability weighting
Probability

ABSTRACT

Psychologists and philosophers who pose moral dilemmas to understand moral judgment typically specify outcomes as certain to occur in them. This contrasts with real-life moral decision-making, which is almost always infused with probabilities (e.g., the probability of a given outcome if an action is or is not taken). Seven studies examine sensitivity to the size and location of shifts in probabilities of outcomes that would result from action in moral dilemmas. We find that moral judgments differ between actions that result in an equal increase in probability of harm (equal size), but have different end-states (e.g., an increase in harm probability from 25% to 50% or from 50% to 75%). This deviation from expected value is robust under separate evaluation, and increases when the comparison between shifts is made explicit under simultaneous evaluation. Consistent with the centrality of perceived harm in some models of moral judgment, perceived harm partially mediates sensitivity to location of harm probability shift. Unlike for shifts in harm probabilities, participants are insensitive to the location of shifts in probability of beneficial outcomes. They are also insensitive to the location of shifts in probability of analogous monetary losses and gains, suggesting an asymmetry between harm and benefit in moral reasoning, as well as an asymmetry between moral and monetary decision-making more broadly. Implications for normative philosophical theory and moral psychological theory, as well as practical applications, are discussed.

1. Introduction

Despite being long neglected by philosophers and psychologists alike, the role of probability in moral decision-making is garnering increasing attention (Fleischhut, Meder, & Gigerenzer, 2017; Ryazanov, Knutzen, Rickless, Christenfeld, & Nelkin, 2018; Shenhav & Greene, 2010; Shou & Song, 2017). Such research bridges the gap between the hypothetical scenarios that stipulate certain outcomes, through which moral dilemmas are traditionally studied, and the real-life scenarios that are infused with outcome uncertainty. For instance, instead of asking "should Tom certainly kill one person to certainly save five people?", the studies mentioned above ask questions such as "should Tom certainly kill one person for a 50% chance of saving ten people?" or "should Tom risk a 50% chance of killing two people to certainly save five people?". Prior studies on uncertain moral dilemmas have identified systematic sensitivity to outcome probabilities in moral judgments, both when it comes to probabilistic saving and probabilistic harming.

1.1. The size and location of probability shifts in moral dilemmas

We distinguish between the *size* of a probability shift and the *location* of a probability shift. The *size* of a probability shift concerns how much the probability of a certain outcome increases or decreases when one performs an action. But a probabilistic shift with the same size—e.g., a 25% increase—can occur in different *locations*, despite resulting in an identical change in expected value. For example, consider a plan that increases the probability of killing four people by 25% in order to save two others. The 25% increase in the probability of four people dying results in an expected loss of one life (0.25×4), whereas certainly saving two people is an expected gain of two lives, leading to a favorable ratio of expected lives saved to expected lives lost (expected value ratio of 2). Note that the 25% increase in probability could occur anywhere in the 0–100% probability interval and not affect expected value. For example, increasing the probability of four people dying from 0% to 25% has the same expected loss of life as does increasing the probability of four people dying from 75% to 100%.

* Corresponding author.
E-mail address: aryazano@ucsd.edu (A.A. Ryazanov).

https://doi.org/10.1016/j.cognition.2020.104548
Received 2 September 2019; Received in revised form 6 December 2020; Accepted 8 December 2020
Available online 25 February 2021
0010-0277/© 2020 Elsevier B.V. All rights reserved.

Thus far, studies on moral dilemmas with uncertainty have all focused on the *size* of probability shifts, and, more specifically, on how the size of probability shifts contributes to the differences in expected value calculation where such shifts are presumed to be increases from 0% or decreases from 100% probability of an outcome. However, the *location* of probability shifts could also matter for moral judgments. In the examples above, it might matter that increasing the probability of a group of four people dying from 0% to 25% means putting at risk a group that is otherwise facing no risk. Or it might matter that the increase from 75% to 100% means that everyone in the group will certainly die. If it turns out that the location of probability shifts has a robust effect on moral judgment *in addition to* the effect from the size of probability shifts, then it becomes puzzling how to characterize the nature of folk moral psychology. The finding would contradict the view that people's moral judgments are *consequentialist*, since being a consequentialist in the traditional sense is to recognize as morally relevant only the difference in expected value of an action, which is independent of the location of the probability shift. Though sensitivity to the location of probability shifts could be consistent with claiming that people's judgments are *deontological*, traditional deontological theories don't have the ready resources to explain the variations either (existing philosophical debates focus on whether contractualist theories, such as Scanlon, 1998, have the resources to justify commonsense moral views about risk-imposition: see Ashford, 2003, Fried, 2012, James, 2012, Kumar, 2015, Frick, 2015, and, for a contractualist autonomy-based defense of a right against risking, Oberdiek, 2017). Moreover, it is unclear why deontological constraints—e.g., that we should not violate people's rights or use people as mere means—are sensitive to *where* the probabilistic shifts occur. Moral dilemmas with different locations of probability shifts are thus worth systematic empirical investigation because of their direct relevance to moral theory and decision-making. In addition, the ways in which people respond to such dilemmas likely have further implications in the areas of actual moral behavior, relations between moral behaviors of different types (such as moral consistency and licensing), and the relation between moral principles and moral behaviors, among others (see Bartels, Bauman, Cushman, Pizarro, & McGraw, 2015). While we leave these future implications for future study, we briefly return to these kinds of implications in the general discussion.

1.2. Relevant research and predictions

Outside of the moral domain, there is evidence suggesting the relevance of the location of probability shifts in decision making. Indeed, Prospect Theory (Kahneman & Tversky, 1979), the most influential theory of decision making under risk, suggests that changes in probability have a nonlinear influence on choices. In particular, people are sensitive to changes near 0% and 100% and relatively insensitive to changes near 50% (Tversky & Kahneman, 1992). For example, Gonzalez and Wu (1999) asked participants to select which of the following felt like a more significant change: increasing the odds of a lottery ticket that has a 65% probability of winning to 70%, or increasing the odds from 90% to 95%. Most participants chose the latter option. When offered a similar choice between increasing the odds of a ticket that has a 5% probability of winning to 10%, or from 30% to 35%, participants were more likely to choose the former option. Similar results have been found when participants make actual choices between monetary gambles (e.g., Abdellaoui, 2000). It remains an open question whether participants would be similarly sensitive to probability shifts in moral dilemmas rather than monetary gambles, and we address this question in the current project. We predict that there will be some kind of sensitivity to the location of probability shifts in moral dilemmas, given the sensitivity to location for monetary gambles.

There may, however, be a difference between the moral and monetary domains in terms of sensitivity to change in probabilities for positive vs. negative outcomes. In the monetary domain, the sensitivity (or discriminability) is the same (or very similar) for gambles involving winning money as it is for gambles involving losing money (Abdellaoui, 2000; Fehr-Duda, De Gennaro, & Schubert, 2006; Pachur & Kellen, 2013; Tversky & Kahneman, 1992). That is, changes in probability near 0% and 100% have a larger effect on choices relative to changes near 50%, regardless of whether gains or losses are involved.[1] But in moral dilemmas, which contain both negative outcomes (harm) and positive outcomes (benefit), there could be a harm/benefit asymmetry in sensitivity to the location of probability shifts, just as there is a harm/benefit asymmetry in moral permissibility judgments (see, e.g., Foot, 1978; Thomson, 1990, chapter 5). For example, according to influential non-consequentialist principles, it is harder, all things equal, to justify killing someone than it is to justify not saving someone, and one explanation is that everyone has a right against everyone else not to be killed but not a right against everyone else to be saved. This asymmetry might be expected to transpose to cases of probabilistic harming and probabilistic saving. In the absence of other morally relevant factors, it may be that there is a *right* to not have one's probability of dying increased whereas there is no right to have one's probability of survival increased. More generally, it is possible that there is a right to not be *probabilistically harmed* (just as there is a right to not be harmed), whereas there is no right to be *probabilistically benefited* (just as there is no right to be benefited).

Empirical evidence adds further support to the philosophical thesis of a harm/benefit asymmetry, which could possibly extend to an asymmetry in sensitivity to harm and benefit probability location shifts. Guglielmo and Malle (2019) find that blame is more differentiated than praise. In particular, they find that mental states preceding negative actions are more finely-distinguished than mental states preceding positive actions (also see Monroe & Malle, 2019). Likewise, negative events have a larger range of linguistic representation than positive events (Peeters, 1971; Rozin & Royzman, 2001). In addition, some empirically-based theories of moral judgment give a much more central role to harm than they do to benefit (e.g., Schein & Gray, 2018). If such discernment extends to sensitivity to various ways of expressing the same change in expected value, but with variation in the location of the probability shift, it suggests that participants may discern more among different locations in probabilistic harm (e.g., whether probability of harm is raised from 0% to 25%, or from 75% to 100%) than among different locations in probabilistic benefit (e.g., whether the probablity of benefit is increased from 0% to 25%, or from 75% to 100%).

We examine whether moral judgments are more sensitive to the location of probability shifts in the case of harm than in the case of benefit, as would be possible if the blame/praise differentiation asymmetry extended to location sensitivity for where shifts in probabilistic harm and benefit occur, and would also be consistent with the asymmetrical treatment of harm and benefit in some major moral theories. We hypothesize that sensitivity to different locations of probability shifts in harm could be attributable to differences in perceived harm. If a shift of harm probability feels like a more significant change, it could be that the action feels more harmful. Though there are presumably deeper explanations for the differences in harm perception, we take it that perceived harm is a good starting point and already goes some way towards giving a more complete story of folk moral psychology in uncertain moral dilemmas.

Although there is a widely espoused moral view about the equal dignity or equal basic moral worth of all persons, according to which there can be no diminishing marginal value of human life (Dworkin, 2002), research on psychological numbing suggests that people are more

[1] The probability weighting function for gains vs. losses has been found to differ in terms of elevation, which corresponds to the impact a given probability has on choices, but not in terms of curvature, which corresponds to discriminability and is our focus in this article (Abdellaoui, 2000; Fehr-Duda et al., 2006; Pachur & Kellen, 2013).

attuned to the suffering of one than the suffering of many (e.g., Dickert, Västfjäll, Kleber, & Slovic, 2015; Kogut & Ritov, 2005; Slovic, 2007, 2010; Västfjäll, Slovic, Mayorga, & Peters, 2014). Such research could be extended to make the prediction that participants will be insensitive to the location of probability shifts involving groups of individuals, as they are, more generally, decreasingly sensitive to helping as the number of victims increases. However, findings on sensitivity to probability in moral decisions involving groups provide evidence against this hypothesis (e.g., Shenhav & Greene, 2010). In addition to exploring sensitivity to location of probability shift, we examine whether psychic numbing could explain any observed sensitivity.

Finally, moral judgments can be studied under separate evaluation, by posing different scenarios to different participants, or under simultaneous evaluation, by asking participants whether they find a particular difference to be morally relevant. Both approaches are useful. Participants provide more consistent moral judgments under simultaneous rather than separate evaluation when presented with two versions of trolley problems—where a person can be dropped onto a trolley track to save others ahead or where a trolley can be diverted onto a track with one person on it to save others ahead (Barak-Corren, Tsay, Cushman, & Bazerman, 2018). One reasonable explanation for this difference is that participants are motivated to reflect on whether their divergent reactions to the two scenarios are normatively defensible under simultaneous evaluation. Thus, we explore the effect of the location of probability shifts in moral judgments under both separate and simultaneous evaluation, under the assumption that separate evaluations will reveal people's unreflective preferences, while simultaneous evaluations will reveal people's considered preferences once the differences in individual cases are made particularly salient.

1.3. Studies

We report seven studies that examine the role of probability shifts in moral judgment. Study 1 finds that participants are sensitive to the location of probability shifts for harm, but not for benefit, when the size of probability shifts is held fixed. Study 2 examines whether the effect of location shifts in harm probability reflects sensitivity to end-state probability and insensitivity to the size of the shift. Studies 3a and 3b explore whether participants endorse the patterns observed under separate evaluation upon reflection, under simultaneous evaluation. Through mediation analyses, Study 4a examines the relationship between location of shift in probability of harming bystanders and perceived harmfulness, while Study 4b examines the relationship between location of shift in probability of saving a group and perceived benefit. Study 5 examines whether sensitivity to location of probability shift differs between analagous moral and monetary decisions.

2. Study 1

We begin by exploring whether participants are sensitive to the location of probability shifts for both harmful and beneficial outcomes in moral dilemmas, when the size of probability shifts is held fixed.

2.1. Study 1 material and methods

One thousand nineteen participants were recruited via Amazon's Mechanical Turk (862 passed an attention check; because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis; 61.8% female; mean age = 34.4, SD = 10.7). Participants were randomly assigned to read one of eight scenarios. Four of the scenarios concerned moral dilemmas in which two people could certainly be saved by increasing the risk of harming four bystanders by 25%. The four scenarios all had different starting and ending points for the probability shift, but the same expected value: 0% to 25%, 25% to 50%, 50% to 75%, and 75% to 100%, see Table 1. As an example, one of the scenarios was as follows:

Table 1
Study 1 scenarios.

Probability shift	Harm Scenarios	Save Scenarios	EV ratio of action
0% to 25%	Increase probability of 4 people dying from 0% to 25% to save 2 people	Kill 1 person to decrease probability of 8 people dying from 25% to 0%	2
25% to 50%	Increase probability of 4 people dying from 25% to 50% to save 2 people	Kill 1 person to decrease probability of 8 people dying from 50% to 25%	2
50% to 75%	Increase probability of 4 people dying from 50% to 75% to save 2 people	Kill 1 person to decrease probability of 8 people dying from 75% to 50%	2
75% to 100%	Increase probability of 4 people dying from 75% to 100% to save 2 people	Kill 1 person to decrease probability of 8 people dying from 100% to 75%	2

> Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There is a group of four bystanders that is facing a 0% risk of death. Harry can carry out a plan that will certainly save the group of two people. However, in carrying out the plan, Harry will increase the risk of death for the group of four bystanders from 0% to 25%.

Participants responded to a single question regarding their confidence that the action should be carried out, adapted to each scenario, along an eleven-point scale. For example, participants were asked, *Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 25%?* (−5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*). We asked about raising the risk of death, rather than imposing a probability of death, in order to describe the probability shifts in language more natural to participants.

The other four scenarios concerned moral dilemmas in which the probability of a group of eight people dying can be decreased by 25% as a result of certainly killing one bystander. Again, four scenarios all had different starting and ending points for the probability shift, but the same expected value: 25% to 0%, 50% to 25%, 75% to 50%, and 100% to 75%. For example, one of the scenarios was as follows:

> Harry sees a group of eight people whose lives are in danger. He knows the following facts. There is a 25% chance of the group of eight people dying. Harry can carry out a plan that will reduce the risk of the group of eight dying from 25% to 0%. However, in carrying out the plan, Harry will certainly kill one bystander.

Participants responded to a single question regarding their confidence that the action should be carried out, adapted to each scenario, along an eleven-point scale. For example, participants were asked, *Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 25% to 0% and at the same time kill one bystander?* (−5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*). We asked about reducing the risk of death, rather than decreasing the probability of death, in order to describe the probability shifts in language more natural to participants.

2.2. Study 1 results

We first verified that participants were generally endorsing an action that had a positive expected value ratio (good done to harm done) of 2. Across all eight versions of the action, testing against the midpoint (0) revealed that participants in general endorsed the action, $t(1018) = 7.01$, $p < .001$, $d = 0.22$ (mean = 0.67, SD = 3.04). There was no significant difference between ratings assigned to scenarios where harm

was probabilistic and scenarios where saving was probabilistic, $t(1017) = 1.11$, $p = .27$, $d = 0.07$ (mean harm = 0.77, $SD = 2.97$; mean save = 0.56, $SD = 3.11$).

Entering where the shift in harming/saving occurred as a linear factor (25–0, 50–25, 75–50, 100–75, recoded as 1, 2, 3, 4, respectively) and entering the type of probability shift (harm or save), as well as their interaction, as factors into an ANOVA revealed an interaction between type of shift and sensitivity to probability shift, $F(1, 1015) = 27.1$, $p < .001$, $r = 0.18$, showing that participants were differentially sensitive to location of probability shifts for probabilistic harming and probabilistic saving (mean harm 0%–25% = 1.55, $SD = 2.54$; mean harm 25%–50% = 1.19, $SD = 2.80$; mean harm 50%–75% = 0.89, $SD = 2.81$; mean harm 75%–100% = −0.55, $SD = 3.29$; mean save 25%–0% = 0.35, $SD = 3.20$; mean save 50%–25% = 0.33, $SD = 2.91$; mean save 75%–50% = 0.76, SD = 3.13; mean save 100%–75% = 0.81, $SD = 3.20$), see Fig. 1. Reverse coding shift location for save shifts (so that 1 = 100–75, 2 = 75–50, etc.) also yielded a significant interaction, $F(1, 1015) = 8.16$, $p = .004$, $r = 0.09$, indicating that the interaction was not an artifact of the coding scheme used to compare harm shifts and saving shifts.

We then separately examined sensitivity to probability shifts on the harm side—is a plan that raises the probability of four bystanders dying from, for example, 0% to 25% in order to save two people preferred to a plan that raises the probability of four bystanders dying from 50% to 75% to save the same number of people? Entering where the shift in harm occurs as a linear factor revealed that participant judgments of whether the action should be carried out were influenced by where the 25% shift in probability of harm occurs, $F(1, 510) = 33.4$, $p < .001$, $r = 0.25$. In order to examine whether this linear effect was driven solely by an aversion to causing certain death—the 75%–100% shift, the linear model was rerun excluding the 75%–100% shift and still yielded a significant linear relationship between decreased confidence in action and the location of the probability shift, $F(1, 510) = 33.4$, $p < .001$, $r = 0.25$. Thus, as the 25% increase in probability of harm occurred closer to 100%, participant confidence in carrying out the action decreased.

We next examined sensitivity to probability shifts on the saving side—for example, is a plan that will certainly kill one bystander in order to reduce the probability of a group of 8 dying from 25% to 0% viewed more favorably than a plan that will certainly kill one bystander in order to reduce the probability of a different group of 8 dying from 75% to 50%? Entering where the shift in saving occurs as a linear factor revealed no significant effect of where the 25% shift in probability of saving occurred on participants' judgments, $F(1, 506) = 2.14$, $p = .14$, $r = 0.065$. To verify that the observed lack of effect on the saving side was not the result of framing the saving probability as a decrease in the probability of dying, rather than an increase in the probability of survival, a separate study compared the two versions for several probability shifts and found no significant sensitivity to probability shift for either frame (see Supplementary Study 1.1). To verify that insensitivity to saving probability shifts was not the result of psychic numbing, or the greater sensitivity to the suffering of the one bystander than the larger group (e.g., Slovic, 2010), several conditions of Study 1 were rerun with the number of individuals involved doubled, such that there was no longer an individual victim—two bystanders could be killed in order to reduce the probability of a different group of sixteen dying by 25% (see Supplementary Study 1.2). We continued to observe an insensitivity to where the shift occurred when accounting for any potential effect of there being an individual victim.

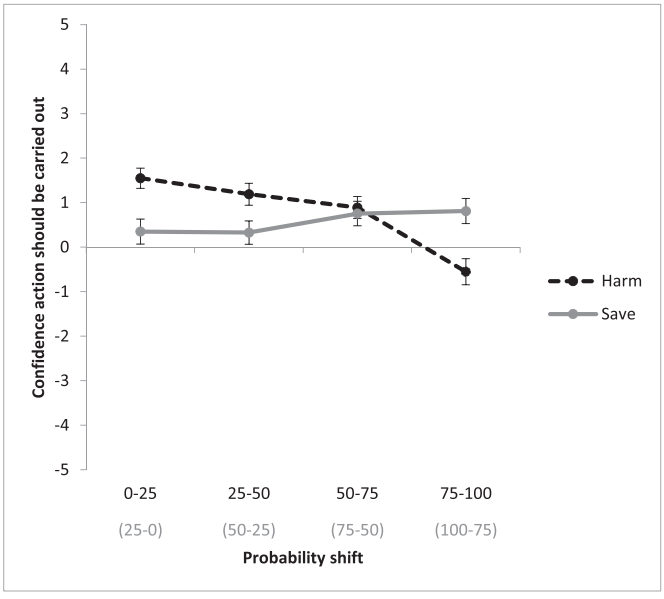

Fig. 1. Sensitivity to probability shift location for an action that increases the probability of four bystanders dying by 25%, in order to save two people (harm), and to probability shift location for an action that decreases the probability of eight individuals dying by 25% by killing one bystander (save). Error bars represent one standard error.

2.3. Study 1 discussion

Study 1 found that that the location of probability shifts affects moral reasoning independently of the ways in which the numerical value of the shift contributes to expected value calculation for probabilistic harm. Harry's plan in each of the eight scenarios concerns exactly the same expected value—the equivalent of 2 lives saved and 1 life lost. Though participants on the whole endorsed action, and we did not observe overall differences in endorsing actions that involved probabilistic harm and actions that involved probabilistic benefit, participants were sensitive to location of the shift in harm probability, but not to the location of the shift in benefit probability.

Thus, there appears to be a harm/benefit asymmetry in sensitivity to the location of probability shifts, when the size of the shift is held fixed. Participants are keenly sensitive to where the shift in probability of harm occurs. However, participants are largely insensitive to where the shift in probability of saving occurs. The finding of differential sensitivity between harm and benefit location shifts is consistent with the more general phenomenon that judgments of negative actions are more fine-grained than moral judgments of beneficial actions (Guglielmo & Malle, 2019; Monroe & Malle, 2019).

The effect of location of harm probability shift appears to increase monotonically, rather than showing increased sensitivity when the shifts occur close to 0% and 100% and decreased sensitivity around 50%, as found by others for monetary decisions (Abdellaoui, 2000; Fehr-Duda, de Genarro, & Schubert, 2006; Gonzalez & Wu, 1999; Pachur & Kellen, 2013; Tversky & Kahneman, 1992). Furthermore, the sensitivity to harm probability shift location does not appear to simply reflect a specific aversion to certain harm, given that excluding the 75%–100% shift from the linear model still yielded a significant linear relationship between decreased confidence in action and the location of the probability shift.

It is not obvious what the sensitivity to the location of probability shifts shows about the nature of folk moral psychology. A common view of moral judgment depicts it as either *consequentialist* or *deontological*. Consequentialist judgments reflect an exclusive concern for the expected values in outcomes, whereas deontological judgments reflect a further concern regarding certain deontological constraints, e.g., constraints against right violation or using people merely as means. But neither consequentialism nor deontology can easily characterize the sensitivity observed in the current study. On one hand, the relevant moral judgments cannot be said to be consequentialist, since to be a consequentialist in the traditional sense is to recognize as morally salient only differences in expected values irrespective of where the probability shifts occur. Thus, the studies clearly reveal that something other than expected value is operating for probabilistic harms. On the other hand, it is not obvious what tools deontologists have to accommodate the

relevance of probability shifts, since it is not clear how the legitimacy of any deontological constraints, e.g., those regarding right violation or treating people as ends, depends on where a probability shift occurs. Thus, the sensitivity observed in the current study constitutes a puzzling phenomenon that cannot be easily accommodated in the traditional consequentialism/deontology framework, though we make some suggestions about how to address this puzzle in the general discussion. This pattern of results may, however, be accounted for psychologically: it could be that shifts in probability of harm that occur closer to certainty feel more harm-like, and that this feeling drives moral judgment. We test this possibility in Study 4, after first examining whether participants attend to both start-state and end-state probabilities, and comparing judgments made under separate and simultaneous evaluation.

3. Study 2

In Study 1, we held fixed the size of probability shifts and explored the role of the location of probability shifts. A further question concerns what will happen if the size of probability shifts is *not* held fixed. Will the end-state of probability shifts affect moral psychology in a way that *overrides* the differences in size and differing starting point? For example, we found that participants were less willing to endorse an action that increased the probability of harm from 75% to 100% than one that increased probability of harm from 50% to 75%. If end-state is all that matters, an increase from 50% to 100% will have the same effect as an increase from 75% to 100%, even though the expected loss of life in the former case is twice that in the latter. Studies 2 and 3 aim to explore this general question from different perspectives.

Study 2 independently varied the size of the probability shift and the end-state of the probability shift for probabilistic harm only. It could be the case that the end-state probability was driving the effect in Study 1. In the event that end-states influence moral judgment, we also wanted to verify that our participants paid adequate attention to both the start-states and the end-states (i.e., the size) of the probability shifts when responding to the moral dilemmas we developed, in order to rule out the possibility that our findings in Study 1 were due to participants not paying attention to initial-state probabilities.

3.1. Study 2 material and methods

Three hundred twenty-nine participants were recruited via Amazon's Mechanical Turk (59.6% female; mean age = 35.2, SD = 10.7). Participants were randomly assigned to read one of four scenarios in which two people could certainly be saved by increasing the probability of death for four bystanders. In two of the scenarios, the shifts were from 0% to 50% and from 25% to 50%; in the other two scenarios, the shifts were from 50% to 100% and from 75% to 100%, see Table 2. Participants then reported their confidence that the action should be carried out on an eleven-point scale, as in Study 1. Immediately after the question, participants were asked to recall the initial probability of harm to the four bystanders and the probability of harm to the four bystanders should the plan be carried out (*What was the initial risk of death for the group of four bystanders?; What would the risk of death be for the group of four bystanders if Harry carried out his plan?*). This was used to verify that participants paid adequate attention to the initial, not just final, probability of harm.

Table 2
Study 2 scenarios.

Probability shift	Harm Scenarios	EV ratio of action
0% to 50%	Increase probability of 4 people dying from 0% to 50% to save 2 people	1
25% to 50%	Increase probability of 4 people dying from 25% to 50% to save 2 people	2
50% to 100%	Increase probability of 4 people dying from 50% to 100% to save 2 people	1
75% to 100%	Increase probability of 4 people dying from 75% to 100% to save 2 people	2

If participants were pure consequentialists, they would more likely endorse actions that involve 25% shifts in probability of harm (EV ratio = 2; 2 lives saved / 1 ended), compared to actions that involve 50% shifts in probability of harm (EV ratio = 1; 2 lives saved / 2 ended). We would expect to see the main effect of shift size, but no effect of end-state, since only shift size matters to EV. However, if participants were instead sensitive to end-state rather than to size of shift, we would expect to see similar ratings of actions with the same end-states (e.g., 100%), regardless of shift size (50%–100% vs. 75%–100%).

3.2. Study 2 results

A 2 × 2 ANOVA analyzing the full set of participants, with size of shift and end-state entered as factors, revealed no significant effect of size of shift, $F(1, 324) = 1.95$, $p = .16$, $r = 0.07$ (mean 50% shift = −0.18, $SD = 3.15$; mean 25% shift = 0.26, $SD = 3.06$), suggesting that participants were not sensitive only to expected value, and therefore not making consequentialist judgments. A significant main effect of end-state revealed that participants were sensitive to whether the end-state was a 50% probability of the four bystanders dying, or a 100% probability of the four dying, $F(1, 324) = 59.9$, $p < .001$, $r = 0.39$ (mean 50% end-state = 1.27, $SD = 2.73$; mean 100% end-state = −1.18, $SD = 2.99$), inconsistent with consequentialism. There was no significant interaction of end-state and size of shift, $F(1, 324) = 0.01$, $p = .93$, $r < 0.01$ (Fig. 2,

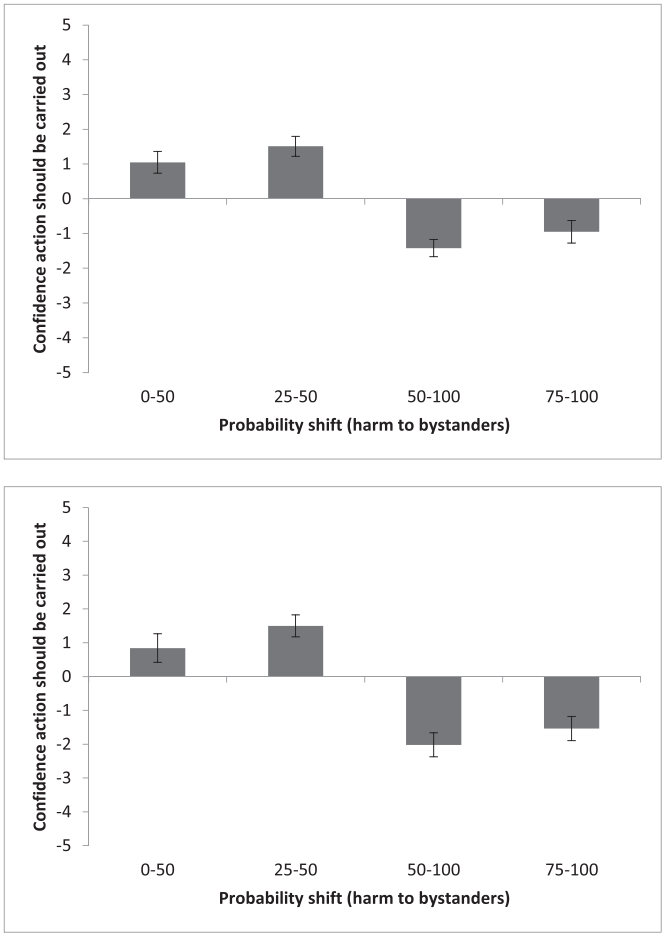

Fig. 2. Sensitivity to size of shift and end-state probability of harm to bystanders for all participants (a) and participants correctly recalling the initial and end-state probability of harm (b). Error bars represent one standard error.

top panel). This suggests two possibilities: participants may have mistakenly attended to only end-states in the experiment, or participants may genuinely care more about end-states than sizes of shift in probability of harm.

By employing a stringent attention check that asked participants to identify starting and ending probabilities using free recall, we could explore whether size of shift mattered for participants who had correctly identified both probabilities. Two hundred thirty-five participants (71%) correctly identified the starting and ending probabilities. Notably, the pattern of results among those participants was similar to the full sample (Fig. 2, bottom panel): A 2 × 2 ANOVA, with size of shift and end-state entered as factors, again revealed no significant effect of size of shift, $F(1,233) = 2.70$, $p = .10$, $r = 0.11$ (mean 50% shift = −0.63, $SD = 3.14$; mean 25% shift = 0.05, $SD = 3.12$). Thus, participants who attended to both start-state and end-state probabilities, and were therefore aware of the shift size, were nonetheless insensitive to shift size. Consistent with the larger sample, we continued to see a significant main effect of end-state, $F(1, 233) = 66.6$, $p < .001$, $r = 0.47$ (mean 50% end-state = 1.19, $SD = 2.84$; mean 100% end-state = −1.77, $SD = 2.71$). There was no significant interaction between end-state and size of shift, $F(1,231) = 0.03$, $p = .86$, $r = 0.01$. Planned contrasts between shifts that resulted in the same end-state, for participants who attended to start-state and end-state probabilities, were all non-significant: 25%–50% vs. 0%–50%: $t(119) = 1.28$, $p = .20$, $d = 0.31$ (mean 25%–50% = 1.45, $SD = 2.70$; mean 0%–50% = 0.84, $SD = 3.00$); 50–100 vs. 75–100: $t(113) = 0.92$, $p = .36$, $d = 0.18$, (mean 50%–100% = −2.02, $SD = 2.60$; mean 75%–100% = −1.54, $SD = 2.81$). Thus, our findings suggest that end-state probabilities, rather than the difference in the sizes of probability shifts or start-states, matter to participants, and that this is not the result of having only attended to end-states.

3.3. Study 2 discussion

Participants were more sensitive to the end-state probability than to the size of the shift in probability of harm to the bystanders, despite being able to recall initial and final probabilities for the scenarios they read. This suggests that participants care more about the final level of probability of harm to the bystanders than about how much the probability of harm has increased.

The fact that end-states matter more than shift size raises the question of just how much people are willing to trade off the two. Large increases in probability of harm with relatively low end-states (e.g., 0%–90%) might be viewed as more acceptable than small increases in probability of harm with high end-states (e.g., 90%–100%). And, if so, we were also interested in whether participants' more reflective preferences under simultaneous evaluation would be consistent with their preferences under separate evaluation.

4. Studies 3a and 3b

Studies 3a and 3b continued exploring how much the location of probability shifts matters. To examine the extent to which participants would prefer a larger increase in probability of harm with a lower end-state probability to a smaller increase in probability of harm with a higher end-state probability, we set out to estimate a balance point value, X, such that participants would no longer prefer the 0% to X% plan over the X% to 100% plan. For a consequentialist, X would be 50 (i.e., indifferent between an increase in probability of harm from 0% to 50% and an increase in probability of harm from 50% to 100%). However, the apparent linear increase in sensitivity to the probability of harm suggests that X will be greater than 50. We examined these preferences under separate and simultaneous evaluation, in order to see whether judgments made under separate evaluation would withstand reflection under simultaneous evaluation.

4.1. Study 3a

4.1.1. Study 3a material and methods

Three hundred ninety-five participants were recruited via Amazon's Mechanical Turk (327 passed an attention check; because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis; 55.4% female, mean age = 33.9, $SD = 11.6$). Each participant was presented with both a separate evaluation task and a simultaneous evaluation task. In the separate evaluation task, participants were randomly assigned to read one of eight scenarios. Four of the eight scenarios involved an increase from 0% to X% probability of harm (X = 50, 75, 85, 95), whereas the other four involved an increase from X% to 100% probability of harm (X = 50, 75, 85, 95). For example, the separate evaluation task that involved an increase from 0% to 95% harm probability went as follows:

> Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There is a group of four bystanders that is facing a 0% risk of death. Harry can carry out a plan that will certainly save the group of two people. However, in carrying out the plan, Harry will increase the risk of death for the group of four bystanders from 0% to 95%.

Participants were asked, for example, *Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 95%?* (−5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*). After completing the separate evaluation, participants then read a simultaneous evaluation scenario in which participants had to choose between two plans, labeled as Plan X and Plan Y. One of these two plans was the plan participants had read under separate evaluation, the second was its matched pair, such that participants saw matched 0% to X% and X% to 100% plans (see Table 3 for scenarios and pairings). For example, participants who had rated either the 0% to 95% plan or the 95% to 100% plan under separate evaluation chose between a 0% to 95% plan and 95% to 100% plan in the simultaneous evaluation task, as follows:

> Harry sees a group of two people who will certainly die without intervention. He knows the following facts. There are two groups, A and B, with four bystanders in each group. Group A is facing a 0%

Table 3
Study 3a scenarios.

Probability shift	Harm Scenarios	Scenario Pairing	EV ratio of action
0% to 50%	Increase probability of 4 people dying from 0% to 25% to save 2 people	A	1
50% to 100%	Increase probability of 4 people dying from 50% to 100% to save 2 people	A	1
0% to 75%	Increase probability of 4 people dying from 0% to 75% to save 2 people	B	0.66
75% to 100%	Increase probability of 4 people dying from 75% to 100% to save 2 people	B	2
0% to 85%	Increase probability of 4 people dying from 0% to 85% to save 2 people	C	0.59
85% to 100%	Increase probability of 4 people dying from 85% to 100% to save 2 people	C	3.33
0% to 95%	Increase probability of 4 people dying from 0% to 95% to save 2 people	D	0.53
95% to 100%	Increase probability of 4 people dying from 95% to 100% to save 2 people	D	10

risk of death, and Group B is facing a 95% risk of death. Harry can carry out a plan, Plan X, that will save the group of two people, but raise the risk of death for the A group of four bystanders from 0% to 95%. Alternatively, he can carry out a plan, Plan Y, that will save the group of two people, but raise the risk of death for the B group of four bystanders from 95% to 100%. He only has time to carry out one of his plans.

Participants were then asked to compare the two plans as follows: *Assuming that Harry must carry out one of the two plans, which should he carry out: Plan X, which he knows with certainty will both save the group of two people and at the same time raise the risk of death for the A group of four bystanders from 0% to 95%; or Plan Y, which he knows with certainty will both save the group of two people and at the same time raise the risk of death for the B group of four bystanders from 95% to 100%? (−5 to 5;very confident Harry should carry out Plan X—not at all confident either way—very confident Harry should carry out Plan Y).*

4.1.2. Study 3a results

In this study, we were interested in whether, upon reflection, participants would endorse an X%-100% plan over a 0%-X% plan. We could thus identify the balance point (X) at which people flip from preferring a 0%-X% plan to an X%-100% plan under simultaneous evaluation, and whether this tracks their preferences under separate evaluation.

First, we examined preferences under separate evaluation. We observed a significant interaction between where the balance point was set (X) and whether the shift in harm resulted in a probabilistic or certain harm end-state (0%-X% or X%-100%), $F(1, 391) = 28.4$, $p < .001$, $r^2 = 0.0725$, see Fig. 3. Consistent with our Study 2 findings, although participants endorsed a plan that raises the probability of four bystanders dying from 0% to 50% to save a group of two, they rejected a plan that raises the probability of the four bystanders dying from 50% to 100% to achieve the same result, $t(96) = 3.48$, $p < .001$, $d = 0.70$ (mean 0%–50% = 0.92, $SD = 2.96$; mean 50%–100% = −1.10, $SD = 2.79$). However, participants preferred a plan that increases the probability of harm from 95% to 100% over a plan that shifts the probability of harm to the four from 0% to 95%, $t(98) = 4.55$, $p < .001$, $d = 0.91$ (mean 0%–95% = −1.85, $SD = 2.80$; mean 95%–100% = 0.75, $SD = 2.90$). There was, on average, no detectable preference between a plan that raises the probability of harm from 0% to 75% and a plan that raises the probability of harm from 75% to 100%, $t(101) = 0.04$, $p = .96$, $d = 0.01$ (mean 0%–75% = −0.58, $SD = 2.95$; mean 75%–100% = −0.60, $SD = 2.95$), nor between a plan that raises the probability of harm from 0% to 85% and a plan that raises the probability of harm from 85% to 100%, $t(92) = 0.69$, $p = .49$, $d = 0.14$ (mean 0%–85% = −0.69, $SD = 3.04$; mean 85%–100% = −0.24, $SD = 3.25$). Thus, under separate evaluation, sensitivity to expected value and to end-state were tied when the balance point (X) was set at 75.

Under simultaneous evaluation, we were interested in whether participants would endorse the patterns observed under separate evaluation (e.g., whether participants endorse the idea that raising the probability of harm to four bystanders from 0% to 50% is preferable to raising the probability of harm to four bystanders from 50% to 100%). While a preference for 0%–50% over 50%–100% could be normatively defensible, in that it would not be unreasonable for avoiding certain harm to be a tie-breaking preference between two actions with matched expected values, a preference for 0%–85% over 85%–100% (or indifference between 0%–95% and 95%–100%) is less clearly defensible, in that the expected value of the 0%–85% action is such that it does much more harm than good (EV ratio = 0.59), whereas the 85%–100% action does much more good than harm (EV ratio = 3.33).

Participants endorsed the pattern of results observed under separate evaluation for 0%–50% and 50%–100% plans, by preferring 0%–50% plans under simultaneous evaluation, despite their equivalent expected value, $t(97) = 6.19$, $p < .001$, $d = 0.63$ (difference from 0; mean = −1.89, $SD = 3.01$), see Fig. 4. Despite not having a detectable preference under separate evaluation between a plan that raises the probability of four bystanders dying from 0% to 75% to one that raises the probability of four bystanders dying from 75% to 100%, participants preferred the 0%–75% plan under simultaneous evaluation, $t(102) = 2.78$, $p = .007$, $d = 0.27$ (mean = −0.86, $SD = 3.15$). Likewise, despite not having a detectable preference between plans under separate evaluation, participants preferred the 0%–85% plan over the 85%–100% plan under simultaneous evaluation, $t(93) = 3.15$, $p = .002$, $d = 0.32$ (mean = −1.03, $SD = 3.18$). Perhaps most surprisingly, despite clearly endorsing the 95%–100% plan and rejecting the 0%–95% plan under separate evaluation, under simultaneous evaluation participants did not exhibit a significant preference between the 95%–100% plan, which has an expected value ratio of 10 (causing much benefit for little harm), and the 0%–95% plan, which has an expected value of 0.53 (causing much more harm than benefit), $t(99) = 1.50$, $p = .13$, $d = 0.15$ (mean = −0.49, $SD = 3.28$). Thus, participants were willing to pay a tremendous expected value premium to avoid raising the probability of death to the bystanders to 100%: Sensitivity to expected value and to end-state, under

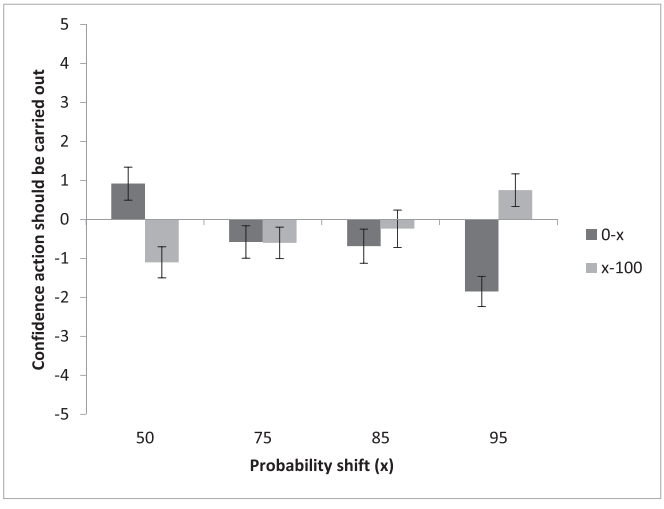

Fig. 3. Sensitivity to location and size of shift in probability of harm to bystanders for plans that increase the probability of death to four bystanders from 0% to X%, or from X% to 100%, in order to save two people, under separate evaluation. Error bars represent one standard error.

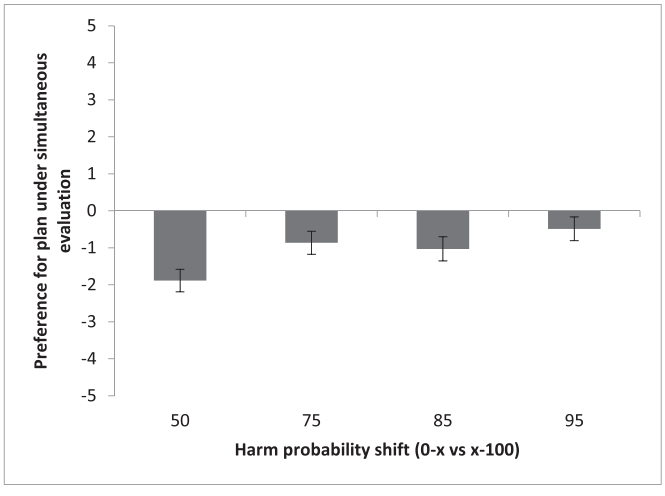

Fig. 4. Preference for plans that increase harm from 0% to X% or X% to 100% to four bystanders in order to save two people, under simultaneous evaluation. Negative numbers indicate a preference for 0%–X% plans. Error bars represent one standard error.

simultaneous evaluation, were tied only once the balance point (X) was set at 95.

One unexpected finding was that an order effect emerged, such that participants who had seen 0%-X% plans under separate evaluation were more likely to endorse them under simultaneous evaluation, $F(1, 387) = 8.76$, $p < .01$ (mean 0%-X% = −1.52, $SD = 3.04$; mean X%-100% = −0.60, $SD = 3.27$), across all scenarios. Supplementary studies were conducted using the X = 50 and X = 95 conditions, where the simultaneous evaluation task was presented without preceding separate evaluations. The results showed the same pattern, with participants being undecided at X = 95, $t(112) = 0.84$, $p = .40$, $d = 0.08$ (mean = −0.26, $SD = 3.24$), and having a strong preference at X = 50, $t(106) = 13.9$, $p < .001$, $d = 2.34$, (mean = −2.89, $SD = 2.15$), see Supplementary Studies 3a.1 and 3a.2.

4.1.3. Study 3a discussion

Under simultaneous evaluation, not only do participants endorse the position that it is better to raise the probability of one dying from 0% to 50% than to raise the probability of another dying from 50% to 100%, but they are willing to pay a premium to avoid raising the probability of harm to the group already at risk. Specifically, they prefer a plan that raises the probability of death for four bystanders from 0% to 85% over a plan that raises the probability of death for another group of four bystanders from 85% to 100%. By contrast, separate evaluation tasks do not show a detectable preference for the 0% to 85% plan. Furthermore, under simultaneous evaluation participants are, on average, indifferent between 0%–95% plans and 95%–100% plans, despite having a strong preference for 95%–100% plans under separate evaluation. We interpret the simultaneous evaluation tasks as revealing people's considered judgments, and they should therefore be taken seriously for the purpose of understanding lay moral theory, as well as for moral theorists' project of engaging in reflective equilibrium in defense of a correct set of moral principles that mutually support and explain our reactions to particular scenarios. But this is subject to further scrutiny. Both the separate and the simultaneous evaluation results add support to our Study 2 finding that the end-states of probability shifts matter more than their size. Surprisingly, simultaneous evaluations (balance point = 95) deviate further from expected value than separate evaluations do (balance point = 75) for shifts in probability of harm. We next examined whether simultaneous evaluation judgments of actions that vary in location of benefit probability shift would be consistent with the separate evaluation insensitivity to location observed in Study 1.

4.2. Study 3b

We have competing predictions regarding what will happen under simultaneous evaluation of scenarios that vary in the location of saving probability shifts. It could be that people continue to be insensitive to the location of probability shifts under simultaneous evaluation, if people genuinely do not believe that it is better to definitely save than to reduce definite risk. Alternatively, it could be that a preference for certain saving emerges under simultaneous evaluation, such that, for example, when choosing between a plan that decreases the probability of death for one group of people from 50% to 0% and a plan that decreases the probability of death for another group from 100% to 50%, people prefer the plan that results in certain saving.

4.2.1. Study 3b material and methods

One hundred seventy-two participants were recruited for this study (141 passed an attention check; because results did not significantly differ between the full sample and those passing the attention check, all participants were retained for analysis, 63.4% female, mean age = 35.3, $SD = 11.0$). Study 3b was structurally similar to Study 3a: Participants were asked to perform a separate evaluation task and a simultaneous evaluation task. But the actions in the scenarios now involved shifts in the probability of a group *dying*, at the cost of killing a bystander. Although we had not observed sensitivity to location of saving probability shift in Study 1, it is possible that a preference for certain saving could emerge under simultaneous evaluation. In order to examine this, we adapted the Study 3a scenarios to describe an action that decreases the probability of a group of eight dying by certainly killing one bystander. We selected new balance point values (X), such that the action could reduce the probability of death for the eight from 100% to X% or from X% to 0%. Selecting X = 50 as a balance allowed us to examine this preference at tied expected values (EV = 4 for both plans). A second balance point, X = 25%, allowed us to determine whether any potential emergent simultaneous evaluation preference for certain saving, observed under tied expected values, could outweigh sensitivity to the expected value of the action when expected values are unmatched (25%–0% EV = 2; 100%–25% EV = 6). Participants were randomly assigned one of four plans: two of the plans involved killing one bystander to decrease the probability of eight dying from X% to 0% (X = 50, 25), and two involved killing one bystander in order to decrease the probability of eight people dying from 100% to X% (X = 50, 25), see Table 4. One scenario, for instance, was as follows:

> Harry sees a group of eight people whose lives are in danger. He knows the following facts. There is a 50% chance of the group of eight people dying. Harry can carry out a plan that will reduce the risk of the group of eight dying from 50% to 0%. However, in carrying out the plan, Harry will certainly kill one bystander.

After evaluating one of the four plans, (e.g., *Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 50% to 0% and at the same time kill one bystander?*; −5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), participants were exposed to the plan they had seen and its matched pair, such that they would see both an X%-0% and corresponding 100%-X% plan, labeled as Plan X and Plan Y:

> Harry sees two groups, A and B, with eight people in each group whose lives are in danger. He knows the following facts. Group A is facing a 50% chance of dying, and Group B is facing a 100% chance of dying. Harry can carry out a plan, Plan X, that will reduce the chance of Group A dying from 50% to 0% but certainly kill one bystander. Alternatively, he can carry out a plan, Plan Y, that will reduce the chance of Group B dying from 100% to 50% but certainly kill one bystander. He only has time to carry out one of his plans.

Participants were then asked to compare the two plans as follows: *Assuming that Harry must carry out one of the two plans, which should he carry out: Plan X, which he knows with certainty will both reduce the risk of the A group of eight dying from 50% to 0% and at the same time kill one bystander; or Plan Y, which he knows with certainty will both reduce the risk of the B group of eight dying from 100% to 50% and at the same time kill one bystander?* (−5 to 5; *very confident Harry should carry out Plan X—not at all confident either way—very confident Harry should carry out Plan Y*).

Table 4
Study 3b scenarios.

Probability shift	Save Scenarios	Scenario Pairing	EV ratio of action
50% to 0%	Kill 1 bystander to decrease probability of 8 people dying from 50% to 0%	A	4
100% to 50%	Kill 1 bystander to decrease probability of 8 people dying from 100% to 50%	A	4
25% to 0%	Kill 1 bystander to decrease probability of 8 people dying from 25% to 0%	B	2
100% to 25%	Kill 1 bystander to decrease probability of 8 people dying from 100% to 25%	B	6

4.2.2. Study 3b results

A 2 (probabilistic vs. certain saving) x 2 (balance point) between-subjects ANOVA was conducted on confidence judgments under separate evaluation. There was a main effect of probabilistic vs. certain saving: Plans that reduced the probability of the eight dying from 100% to X% were rated more favorably than plans that reduced the probability of the eight dying from X% to 0%, $F(1, 170) = 5.74$, $p = .018$, $r = 0.15$ (mean 100%–X% = 1.77, $SD = 2.46$; mean X%–0% = 0.71, $SD = 3.19$), see Fig. 5. There was no main effect of where the balance point was set (whether X = 25 or 50), $F(1, 170) = 0.30$, $p = .58$, $r = 0.04$ (mean 25 = 1.08, $SD = 2.88$; mean 50 = 1.33, $SD = 2.95$), or interaction between where the balance point was set and whether the probability shift resulted in probabilistic or certain saving (100%–X% or X%–0%), $F(1,168) = 0.62$, $p = .43$, $r = 0.06$. The significant main effect of probabilistic vs. certain saving, however, was driven primarily by the 25%–0% vs. 100%–25% comparison, $t(82) = 2.26$, $p = .026$, $d = 0.50$ (mean 25%–0% = 0.47, $SD = 2.53$; mean 100%–25% = 1.87, $SD = 2.53$) rather than the 50%–0% vs 100%–50% comparison $t(86) = 1.12$, $p = .27$, $d = 0.24$ (mean 50%–0% = 0.98, $SD = 3.39$; mean 100%–50% = 1.68; $SD = 2.43$). These results suggest that, under separate evaluation, participants were indifferent between plans with identical expected values but that varied in probability shift end-state (50%–0% and 100%–50%). Participants, were, however, sensitive to the differences in expected value between the 100%–25% (EV = 6) plan and the 25%–0% plans (EV = 2), preferring the plan that resulted in probabilistic saving over the plan that resulted in certain saving, but at a lower expected value. Consistent with Study 1 findings, we observed insensitivity to the location of equivalently-valued saving probability shifts. We additionally found sensitivity to size of saving probability shift, which did not appear to compete with a preference for a certain saving end-state.

Under simultaneous evaluation, participants did not exhibit a significant preference between the plan that reduced the probability of eight dying from 50% to 0% and the plan that reduced the probability of eight dying from 100% to 50%, $t(83) = 0.43$, $p = .66$, $d = 0.05$ (mean = −0.15, $SD = 3.17$), consistent with the insensitivity to end-state of saving probability shift observed under separate evaluation, see Fig. 6. Unlike under separate evaluation, however, participants appeared to be indifferent between the 25%–0% and 100%–25% plans, despite the 100%–25% plan having a higher expected value than the 25%–0% plan, $t(83) = 0.97$, $p = .33$, $d = 0.11$ (mean = 0.36, $SD = 3.36$). Thus, under simultaneous evaluation, we again saw no detectable preference for certain saving end-states, and saw a decreased sensitivity to expected value, as compared to judgments made under separate evaluation.

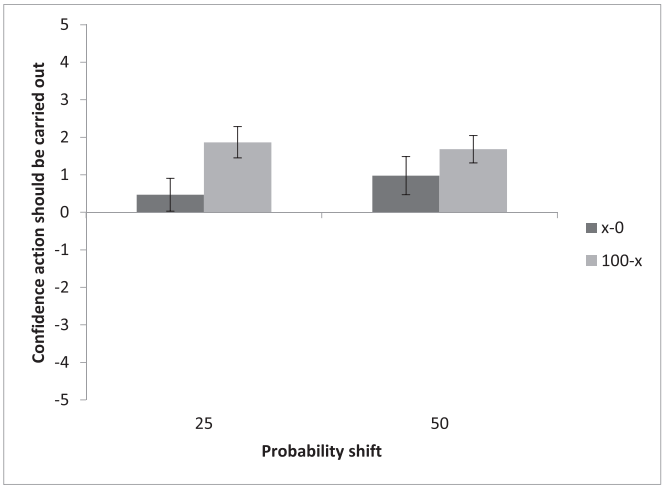

Fig. 5. Sensitivity to location and size of saving probability shift in plans that will reduce the probability of a group of eight dying from X% to 0% or 100% to X% but will kill a bystander, under separate evaluation. Error bars represent one standard error.

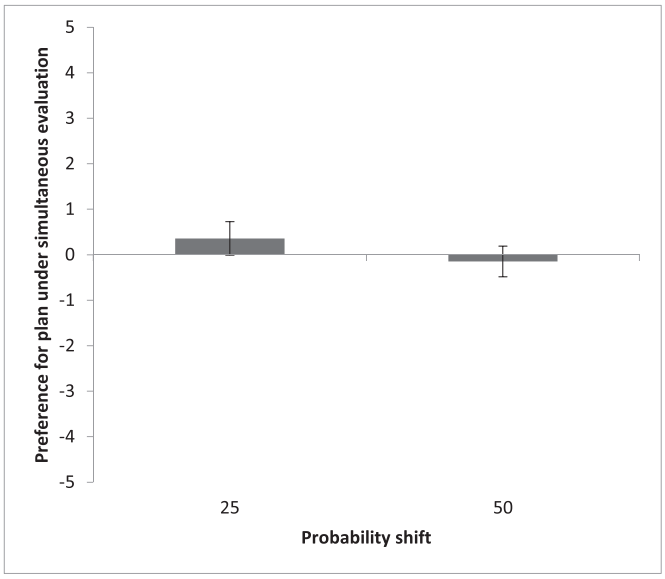

Fig. 6. Preference for plans that reduce the probability of a group of eight dying from X% to 0% or 100% to X% but will kill a bystander, under simultaneous evaluation. Negative numbers indicate a preference for X%–0% plans. Error bars represent one standard error.

4.2.3. Study 3b discussion

Under simultaneous evaluation, participants do not appear to endorse the preference we observed under separate evaluation for a 100%–25% (rather than 25%–0%) shift in probability of the group of eight dying. This seperate evaluation preference can be interpreted as a sensitivity to size of shift, as participants in Study 1 were insensitive to location of saving probability shifts when the size of the shift was held constant. The overall findings, however, are consistent with participants being relatively insensitive to the location of saving probability shifts, as compared to harming probability shifts, as observed in prior studies. Again, here we take the results from simultaneous evaluations to be participants' considered preferences and to reflect their lay moral theory. Though the lack of sensitivity to the location of probabilistic benefit confirms our hypothesis that there is more sensitivity on the side of harm than on the side of benefit, it comes as a surprising finding that, despite the size of shift being relevant under separate evaluation, we observe no sensitivity under simultaneous evaluation for this. Although the result is broadly consistent with existing findings on the harm/benefit asymmetry, according to which moral reasoning may be more fine-grained in sensitivity to shifts in harm than to shifts in benefit, it is still surprising that folk moral psychology is so coarse-grained that it doesn't appear to differentiate between, for example, decreasing the probability of people dying from 100% to 25% and decreasing the probability of the same number of people dying from 25% to 0% under simultaneous evaluation.

The general harm vs. benefit asymmetry, as observed in Study 1, and reconfirmed in Studies 3a and 3b, is consistent with the hypothesis we started with: There should be more sensitivity on the harm side, perhaps because we have a right or claim to not have our probability of dying increased, whereas we do not have a right to have our probability of living increased (Oberdiek, 2017).

5. Studies 4a and 4b

Throughout the preceding studies, we observed a sensitivity to where shifts in probability of harm occur and a general insensitivity to where shifts in probability of saving occur. One question that emerges is why such a sensitivity occurs. One possibility is that participants perceive

different harm probability shift locations as differentially harmful. Though participants might focus exclusively on perceived harm, it is plausible that perceived benefit might play a mediating role as well in cases involving probabilistic saving. In the following studies, we explore whether differences in perceived harm and perceived benefit are causally relevant to divergent moral judgments of actions that vary in location of probability shift.

5.1. Study 4a material and methods

Study 4a asked participants to evaluate one of two plans, as in Study 1. The first plan involved increasing the probability of dying to a group of four bystanders from 0% to 25% in order to save two people. The other plan involved a 75% to 100% increase in the probability of the group of four bystanders dying to save the two. After evaluating one of the two plans, participants were asked two questions: one regarding whether Harry should carry out the plan (e.g., *Should Harry carry out a plan that he knows with certainty will both save the group of two people and at the same time raise the risk of death for the group of four bystanders from 0% to 25%?*; −5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), and a second regarding how harmful the action would be to the group of bystanders (e.g., *How harmful is Harry's plan for the four bystanders?;* 0: *not at all harmful*, to 10: *extremely harmful*). Though past researchers have used a three-item measure of perceived harm (how threatening, dangerous, harmful, see Schein, Ritter, & Gray, 2016), we chose to assess perceived harm via harmfulness—the most directly relevant item from this composite measure for our moral dilemmas. Questions were presented in a randomized order. Two hundred twenty-two participants were recruited using Amazon's Mechanical Turk (163 passed an attention check; because results do not significantly differ between full sample and those passing attention check, all participants were retained for analysis; 54.1% female, mean age = 35.1, *SD* = 11.5).

5.2. Study 4a results

To examine whether perceived harm to the bystanders mediates moral judgment of actions that vary in location of harm probability shift, we first tested the significance of each individual path. The relationship between harm probability shift (0%–25% or 75%–100%) and confidence in the morality of action was partially mediated by the perceived harmfulness of the action for the group of four bystanders. The regression of probability shift on confidence in the morality of action was statistically significant ($\beta = 0.37$, $t(220) = 6.34$, $p < .001$), as was the regression of probability shift on perceived harmfulness ($\beta = 0.54$, $t(220) = 9.40$, $p < .001$), and the regression of perceived harmfulness on confidence in the morality of action ($\beta = 0.37$, $t(220) = 5.91$, $p < .001$), see Fig. 7. The standardized indirect effect was (0.54)(0.37) = 0.27. The significance of this indirect effect was tested using bootstrapping procedures: Unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect partially mediated the relationship between probability shift and judgments of the action (mediated effect = 1.06, $p < .001$, 95% CI [0.50, 1.66]; direct effect = 2.39, $p < .001$, 95% CI [1.38, 3.43]).

5.3. Study 4a discussion

Perceived harm partially mediated the difference between plans that increase the probability of bystanders dying from 0% to 25% and plans that increase the probability of bystanders dying from 75% to 100%. We next explore whether we see a similar effect for saving probability shifts.

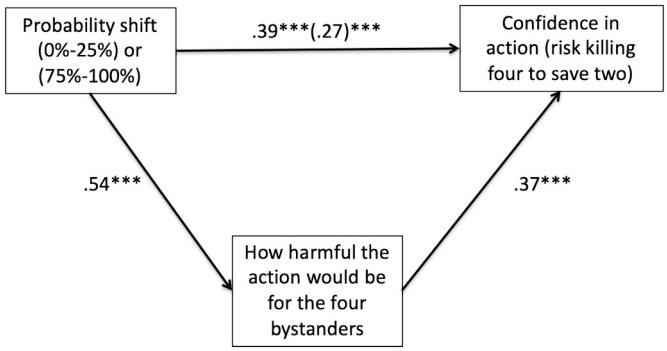

Fig. 7. Standardized regression coefficients for the relationship between harm probability shift and confidence in the morality of carrying out an action that raises the probability of four bystanders dying to save people as mediated by perceived harmfulness of the action for the four bystanders. The standardized regression coefficient between probability shift and confidence in the morality of action, controlling for perceived harmfulness, is in parentheses.
***$p < .001$.

5.4. Study 4b

5.4.1. Study 4b material and methods

Study 4b adopted the approach of Study 4a to scenarios in which the saving is probabilistic, and in which location of shift in saving probability varies. Study 4b asked participants to evaluate one of two plans. The first plan would result in the death of one bystander in order to decrease the probability of a group of eight people dying from either 25% to 0% or 100% to 75%. After evaluating one of the two plans, participants were asked two questions, one regarding whether Harry should carry out the plan (e.g., *Should Harry carry out a plan that he knows with certainty will both reduce the risk of the group of eight dying from 25% to 0% and at the same time kill one bystander?*; −5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*), and a second regarding how beneficial the action would be for the eight people (e.g., *How beneficial is Harry's plan for the eight people whose lives are in danger?;* 0: *not at all beneficial*, to 10: *extremely beneficial*). Questions were presented in a randomized order. Two hundred twenty-five participants were recruited using Amazon's Mechanical Turk (165 passed an attention check; because results do not significantly differ between full sample and those passing attention check, all participants were retained for analysis; 56.9% female, mean age = 32.4, *SD* = 10.7).

5.4.2. Study 4b results

In order for us to test whether perceived benefit mediates confidence in the morality of actions that carry a variy in benefit probabilty shift location, we first tested the significance of each individual path. The regression of probability shift location on confidence in action was not statistically significant ($\beta = 0.07$, $t(223) = 1.08$, $p = .28$). The non-significant effect observed here is consistent with Study 1 findings; however, we can still examine whether perceived benefit varies with saving probability shift location, and exerts an indirect effect on moral judgment. The regression of probability shift on perceived benefit was significant ($\beta = 0.20$, $t(221) = 3.05$, $p < .01$), as was the regression of perceived benefit on confidence in the morality of action ($\beta = 0.28$, $t(221) = 4.32$, $p < .001$). The standardized indirect effect was (0.20)(0.28) = 0.15. The significance of this indirect effect was tested using bootstrapping procedures: unstandardized indirect effects were computed for each of 10,000 bootstrapped samples, and the 95% confidence interval was computed by determining the indirect effects at the 2.5th and 97.5th percentiles. The bootstrapped unstandardized indirect effect partially mediated the relationship between probability shift and judgments of the action (mediated effect = 0.38, $p = .001$, 95% CI [0.12, 0.73]; direct effect = 0.91, $p = .023$, 95% CI [0.12, 1.7]), suggesting that perceived benefit is sensitive to outcome probability shifts, and that

perceived benefit does relate to confidence in the morality of action, though not in a way that contributes to overall differences between the two versions of the scenario, see Fig. 8.

5.4.3. Study 4b discussion

Though perceived benefit does correlate with confidence in the morality of action more generally, and indirectly mediates a difference between moral judgments where the saving probability shift location differs between two otherwise identical moral dilemmas, we did not observe an overall sensitivity to saving probability location, consistent with Study 1 findings.

5.5. Studies 4a-b discussion

Participants differentiated more between probability shift locations for harm than for benefit: The effect size of saving probability shift location on perceived benefit ($\beta = 0.20$) was less than half of the effect size of harm probability shift location on perceived harm ($\beta = 0.53$). This greater differentiation for harm-related probability shift locations is consistent with a more general greater differentiation of negative events than positive events other researchers have observed (e.g., Guglielmo & Malle, 2019), and provides a mechanism for the sensitivity to location of harm probability shifts, but not location of saving probability shifts, observed in earlier studies.

6. Study 5

Thus far, we have observed that participants are sensitive to the location of changes in probability regarding harming, but not saving, individuals, and they are especially sensitive to changes that result in a high probability of harm. This differential sensitivity is inconsistent with other empirical research examining choices between monetary gambles, which has shown no difference in the curvature, or discriminability, of the probability weighting function for gains vs. losses (Abdellaoui, 2000; Fehr-Duda et al., 2006; Pachur & Kellen, 2013). To see whether this differential discriminability is unique to the moral domain, Study 5 dicompared our moral scenarios to analgous monetary decisions. Additionally, to verify that our findings are not driven by a sensitivity to certain end-state (100% or 0%), we adapted the scenarios to involve shifts to (and from) 5% and 95%, rather than shifts to (and from) certainty.

6.1. Study 5 material and methods

We tested sensitivity to probability shifts from 5% to 50% and from 50% to 95% for moral scenarios involving raising risk of harm to ten individuals in order to certainly save five individuals, moral scenarios involving certainly killing four individuals to reduce risk of harm to ten individuals, and monetary loss/gain scenarios written to be analogous to them. Here, for example, is the 5% to 50% loss probability shift monetary scenario:

> Harry knows the following facts about his own investments. If he does nothing, he will certainly lose Investment A, a $5000 investment. He is also facing a 5% risk of losing Investment B, a $10,000 investment. Harry can carry out a plan that will certainly recover Investment A. However, in carrying out the plan, Harry will increase the risk of losing Investment B from 5% to 50%

Participants read one of eight scenarios (moral scenarios involving an increase of 5% to 50% or 50% to 95% increase in harm, monetary scenarios involving an increase of 5% to 50% or 50% to 95% increase in chance of loss, moral scenarios involving a decrease in harm from 50% to 5% or 95% to 50%, monetary scenarios involving a decrease in chance of loss from 50% to 5% or 95% to 50%, see Table 5. For monetary scenarios, participants were asked a question that paralleled the moral question, e.g., *Should Harry carry out a plan that he knows with certainty will both save Investment A, a $5000 investment, and at the same time raise the risk of losing Investment B, a $10,000 investment, from 5% to 50%?* (−5: *very confident Harry should not carry out the plan*, to 5: *very confident Harry should carry out the plan*). Eight hundred nine participants were recruited using Amazon's Mechanical Turk (709 passed an attention check; because results do not significantly differ between full sample and those passing attention check, all participants were retained for analysis; 55.0% female, mean age = 35.3, $SD = 12.0$).

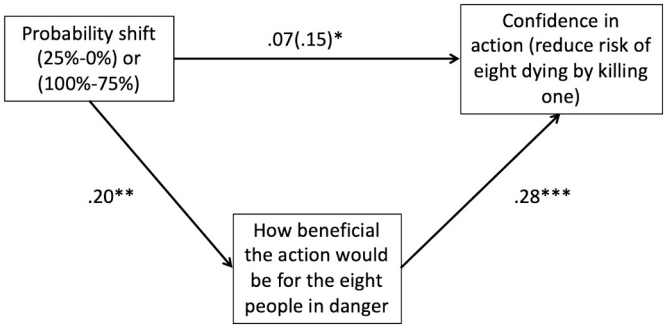

Fig. 8. Standardized regression coefficients for the relationship between benefit probability shift and confidence in the morality of carrying out an action that kills one bystander to reduce the risk of eight people dying as mediated by how beneficial the action is perceived to be for the eight people. The standardized regression coefficient between probability shift and confidence in the morality of action, controlling for perceived benefit, is in parentheses.
$^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

6.2. Study 5 results

We first separately examined sensitivity to location of harm and loss probability shifts for the moral and monetary scenarios, respectively. As in our earlier studies, we observed a sensitivity to location of harm probability shift in moral scenarios, $t(202) = 3.08$, $p = .002$, $d = 0.43$ (moral 5%–50% harm mean = 0.71, $SD = 3.01$ moral 50%–95% harm mean = −0.56, $SD = 2.90$). For analogous monetary scenarios, however, we did not observe sensitivity to location of loss probability shift, $t(197)$

Table 5
Study 5 scenarios.

Probability shift	Scenarios	Scenario Pairing	EV ratio of action
5% to 50% harm shift	Increase probability of 10 people dying from 5% to 50% to save 5 people	Moral	1.11
50% to 95% harm shift	Increase probability of 10 people dying from 50% to 95% to save 5 people	Moral	1.11
5% to 50% loss shift	Increase probability of losing a $10,000 investment from 5% to 50% to save a $5000 investment	Monetary	1.11
50% to 95% loss shift	Increase probability of losing a $10,000 investment from 50% to 95% to save a $5000 investment	Monetary	1.11
50% to 5% saving shift	Kill 4 people to decrease probability of 10 people dying from 50% to 5%	Moral	1.12
95% to 50% saving shift	Kill 4 people to decrease probability of 10 people dying from 95% to 50%	Moral	1.12
50% to 5% gain shift	Lose a $4000 investment to decrease the probability of losing a $10,000 investment from 50% to 5%	Monetary	1.12
95% to 50% gain shift	Lose a $4000 investment to decrease the probability of losing a $10,000 investment from 95% to 50%	Monetary	1.12

= 0.17, p = .87, d = 0.02 (monetary 5%–50% loss mean = 0.12, SD = 2.95, monetary 50%–95% loss mean = 0.051, SD = 2.77). A significant interaction of type of scenario (moral or monetary) and location of shift in harm/loss probability (5%–50% or 50%–95%), $F(1, 399) = 4.34$, p = .038, $\eta^2 = 0.011$, confirmed that sensitivity to probability shift location occured in the moral, but not the monetary, scenario, Fig. 9, top panel. There was no main effect of domain for the harm/loss scenarios, such that, overall, participants were equally willing to endorse the monetary action and moral action $F(1, 399) = 0.003$, p = .95, $\eta^2 < 0.001$. (mean moral = 0.069, SD = 3.02, mean monetary = 0.085, SD = 2.83). There was a main effect of location of probability shift, $F(1, 399) = 5.49$, p = .020, $\eta^2 = 0.013$ (50%–95% shift mean = −0.26, SD = 2.85; 5%–50% shift mean = 0.42, SD = 2.99), driven by the moral scenarios.

Next we examined sensitivity to location of probability shift in saving (expressed as a reduction in risk of harm or loss) for the moral and monetary scenarios. As in our earlier studies, we did not observe a sensitivity to location of saving probability shift in moral scenarios, $t(202) = 0.41$, p = .68, d = 0.06 (moral 50%–5% save mean = −0.54, SD = 3.30; moral 50–95% save mean = −0.72, SD = 2.96). The same was true for our analogous monetary scenarios: There was a lack of sensitivity to location of gain probability shift, $t(200) = 0.61$, p = .54, d = 0.09 (monetary 5%–50% gain mean = 1.04, SD = 2.51; monetary 50–95% gain mean = 0.804, SD = 2.99). A nonsignificant interaction of domain (moral or monetary) and location of probability shift (5%–50% or 50%–95%), $F(1, 402) = 0.0094$, p = .92, $\eta^2 < 0.001$, confirmed consistent insensitivity to probability shift location across monetary gains and moral saving, see Fig. 9, bottom panel. Additionally, there was a main effect of domain for the save/gain scenarios, such that participants were more willing to endorse the monetary action than the moral action, $F(1, 402) = 28.0$, $p < .001$, $\eta^2 = 0.065$. (mean moral = −0.63, SD = 3.12; mean monetary = 0.92, SD = 2.76). There was no main effect of location of probability shift for these scenarios, $F(1, 402) = 0.50$, p = .48, $\eta^2 = 0.001$ (50–95% shift mean = 0.04, SD = 3.06; 5–50% shift mean = 0.24, SD = 3.03).

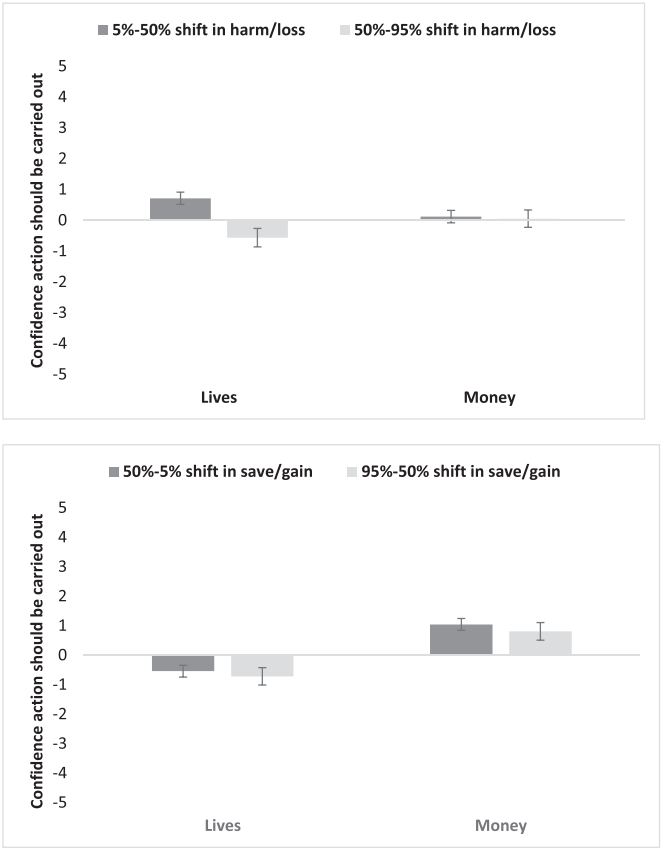

Fig. 9. Confidence in action for plans that increase the chances of harming ten bystanders in order to save five individuals, and that increase the chances of losing a $10,000 investment in order to save a $5000 investment (a). Confidence in action for plans that kill four bystanders in order to reduce the risk of ten individuals dying, and that lose a $4000 investment in order to reduce the risk of losing a $10,000 investment (b). Error bars represent one standard error.

6.3. Study 5 discussion

Our repeated finding of increased sensitivity to changes in probability that are closer to 100% for negative outcomes in moral scenarios appears not to apply to monetary scenarios. There seems to be something special about moral dilemmas in this respect. Moreover, we replicated the moral harm pattern of results using shifts in probability that did not include 0% or 100%, which, along with the results of Study 1, show that it is not just the endpoints (certain harm, in particular) driving our results. We continue to observe insensitivity to probability shifts affecting positive outcomes in moral scenarios, and we also found this for monetary scenarios.

Our finding that there was not greater sensitivity to changes in high (rather than low) probability of monetary gains appears to be at odds with recent findings of Lewis and Simmons (2020). These authors reported that, for example, participants were willing to pay more to increase the probability of winning a prize from 80% to 89% than to increase the probability of winning from 11% to 20%, which could indicate greater sensitivity to change in probability closer 100%. Even when those authors examined scenarios involving avoiding losses – which is more similar to our "gain" scenarios – they still found that participants were willing to pay more for changes to high, rather than low, probabilities. It is difficult to know why they find different results, since there are many differences between their materials and procedures and ours (e.g., they ask for willingness to pay, whereas we ask about confidence in carrying out plans; their scenarios are first person, whereas ours are third person). Perhaps most revealing is that Lewis and Simmons provide evidence that differential sensitivity is not driving their effect, which is instead due to the greater appeal of the higher endpoint (e.g., 89% probability of winning) compared to the lower endpoint (e.g., 20% probability of winning; see their Study 7), suggesting that our studies are tapping into different processes. Further research is required to sort out how the different task characteristics lead to different results.

7. General discussion

Our results suggest that participants are sensitive not only to expected value and probability in moral dilemmas, but also to where the shift in probability occurs, at least for increases of risk to bystanders. Study 1 showed that participants are sensitive to where shifts in the probability of harm to a group of bystanders occur, but not to where shifts in the probability of saving a group occur. Study 2 found that participants are sensitive to end-state probability, rather than to the size of the shift or start-state probability, for shifts in probability of harm to bystanders. Study 3a explored whether participants would endorse such a position upon reflection, under simultaneous evaluation, and found that participants prefer plans that raise the probability of harm to a group of four bystanders from 0% to 85% over plans that raise the probability of harm from 85% to 100% for a different group of four bystanders, in order to save two people. In Study 3b, we found that participants continued being indifferent between locations of shifts in probability of saving a group under simultaneous evaluation. Study 4a identified perceived harm as a mechanism behind divergent moral judgments for actions that cause mathematically equivalent harm probability shifts occurring at different parts of the probability distribution. Study 4b found that, though saving probability shifts are perceived as differentially beneficial, differences in perceived benefit

between probability shifts do not result in divergent moral judgments. Study 5 identified that the observed pattern of sensitivity to location of shift in harm (or loss) probability applies to moral, but not monetary, decisions.

In general, the results from these studies show that participants are strongly sensitive to the location of probability shifts for harm, that locations of probability shifts matter in ways that can make participants indifferent between probability shifts of different sizes, and even that participants prefer a *larger* size of increase of the probability of people dying just because the increase occurs in the preferable location. Thus, the location isn't merely a "tie-breaker" for participants but rather plays a substantial role in their decision-making process.

These phenomena make it difficult to render folk moral judgments consistent with the traditional consequentialist framework which takes expected value to be the single currency for moral permissibility and obligation. At the same time, the traditional tools of the deontologist, which presuppose the existence of moral constraints on certain harming, do not capture this phenomenon without further elaboration. We thus need a more nuanced way to describe the normative principle that can explain why participants reason in this way. We believe that one promising strategy is to develop more fine-grained deontological constraints that concern moral decision-making under uncertainty. Such a strategy would need to, for example, explicate the force of deontological constraints in terms of the location of probability shifts. For example, it may be that the right against raising a person's risk of death to 100% is more stringent than the right against raising a person's risk of death to 50%, even when the probability shift in the latter case is greater. We are optimistic that this kind of framework can be worked out, but a complete defense has to be left for another occasion.

We also see a general harm/benefit asymmetry in sensitivity to locations of probability shifts. This pattern of asymmetry adds further support to the claim that participants are sensitive to probability shifts in harm in a more fine-grained way than they are to probability shifts in benefit. However, the finding that there is no significant variation according to the locations of probability shifts on the benefit side is surprising to us, potentially suggesting a deeper harm/benefit asymmetry than first appears plausible. We speculate that participants believe, or act in ways that are properly explained by the belief, that people have a right or claim not to be harmed, but no right or claim to be benefited. Whether there is explicit additional evidence for this hypothesis is a matter worthy of further investigation.

Future research could examine the extent to which the location of probability shifts affects moral psychology beyond decision-making (Bartels et al., 2015), such as actual moral behavior (e.g., Bostyn, Sevenhant, & Roets, 2018), the relation between moral behaviors (i.e., moral consistency and moral licensing; see Mullen & Monin, 2016), and the relation between moral principles and behavior (e.g., moral hypocrisy; see Monin & Merritt, 2012). For example, given recent research suggesting that moral judgments may not be predictive of moral action (e.g., Bostyn et al., 2018), it is of interest whether our findings predict moral action. Similarly, because the relation between moral behaviors has been studied only in the context of certain outcomes, the extent to which the location of probability shifts can affect moral licensing and consistency effects could be explored. Finally, consistent with the challenge of reconciling traditional philosophical views with the observed sensitivity to the location of harm probability shifts, further research is needed to understand folk theories of how behavior and moral principles are reconciled (e.g. moral hypocrisy) for actions that involve shifts in probabilities of harm.

The more practical applications of the observed pattern of sensitivity to locations of probability shifts can be seen, for example, in the design of autonomous vehicles. Autonomous vehicles all face "decisions" under uncertainty; it is therefore worth examining how complicated the probabilistic computing would need to be, if we want these vehicles to make decisions consistent with folk moral psychology. Our results show that the computation has to involve at least two distinct probability estimates: an estimate of initial (or final) probabilities of various outcomes and an estimate of the size of probability shifts. That is, it is not enough for autonomous vehicles to simply calculate *how much* change in the probability of harming people would be involved in alternative courses of action. One implication of the finding that participants care more about end-state harm probabilities than sizes of shifts is that initial probability levels are largely irrelevant to the endorsement of action in moral dilemmas. Such findings help inform where resources should be directed in the design of detectors in autonomous vehicles that will serve as inputs for ethical decisions these vehicles will be programmed to make: when facing a decision as to whether autonomous vehicle carrying passengers should be redirected into a pedestrian to avoid a fatal collision, that pedestrian's initial probability of harm arising from this collision (e.g., whether inaction would lead to a 25% chance or 75% chance of them dying) is less relevant than the end-state probability of harm to the pedestrian the redirection would cause.

An additional implication regards theories of decision-making under risk and, in particular, people's sensitivity to changes in probability. Previous empirical work (Abdellaoui, 2000; Fehr-Duda et al., 2006; Pachur & Kellen, 2013; Tversky & Kahneman, 1992), using choices between monetary gambles as data, has found that sensitivity to changes in probability (discriminability) is virtually the same for positive and negative outcomes (i.e., gains and losses). While we also found this for our monetary scenarios in Study 5, we found very different sensitivity to location of changes in probability of harm than for changes in probability of saving in all of our moral scenarios. Looking at only negative outcomes, we found greater sensitivity to probability shift location for moral scenarios than for monetary scenarios. Thus, reaction to changes in probability for negative moral outcomes seems different not only from reactions to positive moral outcomes, but from negative monetary outcomes as well. Because monetary gambles are the fruit fly of decision-making research, it is of considerable interest that people treat uncertainty differently in our moral scenarios. Using only choices between monetary gambles to inform models of decision-making under risk may thus limit the scope of the models (see also Müller-Trede, Sher, & McKenzie, 2018).

Recent research by Evers and Imas (2019) suggests that, when outcomes are similar, they are bracketed into a single mental account, and when different, they are not. Monetary gains and losses can function as inputs in calculations related to a single mental account, or can be treated as belonging to different mental accounts. While we did not directly test whether differences between how lives and money are bracketed contributed to our results, our observed results cannot be explained by a simple difference in bracketing. We observed insensitivity to location of probability shift for monetary gains, monetary losses, and lives saved, but not lives lost. If a main effect of how lives and money are bracketed underpinned our results, we would have observed parallel patterns of results for moral gains and moral losses. Furthermore, studies of moral decisions with probabilistic outcomes, such as the Asian Disease Problem, find results similar to studies that use probabilistic monetary outcomes (Tversky & Kahneman, 1981). It is possible that the imposition of harm on bystanders is less easily bracketed than the reduction in risk of harm, but further research is needed to determine whether bracketing contributes in this nuanced way to the observed differences in moral and monetary decision making.

Finally, separate and simultaneous evaluations differed in our studies. Like others who have advocated for simultaneous evaluations to be taken more seriously, since they indicate participants' considered preferences under reflection (Barak-Corren et al., 2018), we find moral judgment under separate and simultaneous evaluation to diverge, calling into question the common method both in philosophy and in psychology to simply test for moral permissibility by appealing to intuitions about single cases. However, we do not find judgments under simultaneous evaluation to be more clearly defensible, though we can say that simultaneous evaluation judgments of actions involving shifts in harm or saving deviate further from expected value calculation than separate

evaluations for some actions. A complete moral theory would benefit from further comparison across both kinds of evaluation to come to an eventual moral verdict. This, again, has crucial implications for designing autonomous vehicles. In order to be consistent with folk moral psychology, the relevant moral principles would be nuanced in a way that reflects whether the vehicle is facing a single alternative course of action to a collision, or multiple options, in addition to the end-state probability for each outcome. More broadly, such research informs decision-making in complicated real-life cases in which there are multiple actions that might be taken, each matched with a variety of probabilities regarding outcomes.

7.1. Conclusions

The results of seven studies suggest that participants are sensitive not only to the expected value of actions that harm some in order to benefit others, but also to the end-state probability of harm that would result from an action that harms some to benefit others. Perceived harm is identified as a mechanism contributing to the divergence in moral judgments of actions that cause mathematically equivalent harm probability shifts, but result in different end-states. Sensitivity to end-state probability appears to be unique to shifts in probability of harm, not to shifts in probability of benefit, and does not occur for analogous monetary decisions. Results suggest that folk moral judgments fit neither traditional consequentialist frameworks, nor traditional deontological frameworks, and have implications for applied ethics.

Disclosure statement

For all experiments, we have reported all measures, conditions, and data exclusions. Sample sizes were determined such that all conditions would have samples consistent with prior literature on probability shifts in moral dilemmas (e.g., Ryazanov et al., 2018).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2020.104548.

References

Abdellaoui, M. (2000). Parameter-free elicitation of utility and probability weighting functions. *Management Science, 46*, 1497–1512.

Ashford, E. (2003). The demandingness of Scanlon's contractualism. *Ethics, 113*(2), 273–302.

Barak-Corren, N., Tsay, C., Cushman, F., & Bazerman, M. (2018). If you're going to do wrong, at least do it right: The surprising effect of considering two moral dilemmas at the same time. *Management Science, 64*(4), 1528–1540.

Bartels, D. M., Bauman, C. W., Cushman, F. A., Pizarro, D. A., & McGraw, A. P. (2015). Moral judgment and decision making. *The Wiley Blackwell handbook of judgment and decision making, 63*, 478–515.

Bostyn, D. H., Sevenhant, S., & Roets, A. (2018). Of mice, men, and trolleys: Hypothetical judgment versus real-life behavior in trolley-style moral dilemmas. *Psychological Science, 29*(7), 1084–1093.

Dickert, S., Västfjäll, D., Kleber, J., & Slovic, P. (2015). Scope insensitivity: The limits of intuitive valuation of human lives in public policy. *Journal of Applied Research in Memory and Cognition, 4*(3), 248–255.

Dworkin, R. (2002). *Sovereign virtue: The theory and practice of equality*. Cambridge, MA: Harvard University Press.

Evers, E., & Imas, A. (2019). Mental accounting, similarity, and preferences over the timing of outcomes. Available at SSRN https://ssrn.com/abstract=3452943. or https://doi.org/10.2139/ssrn.3452943.

Fehr-Duda, H., de Genarro, M., & Schubert, R. (2006). Gender, financial risk, and probability weights. *Theory and Decision, 60*, 283–313.

Fleischhut, N., Meder, B., & Gigerenzer, G. (2017). Moral hindsight. *Experimental Psychology, 64*(2), 110–123.

Foot, P. (1978). *Virtues and vices*. Oxford: Blackwell.

Frick, J. (2015). Contractualism and social risk. *Philosophy and Public Affairs, 43*(3), 175–223.

Fried, B. H. (2012). Can contractualism save us from aggregation? *The Journal of Ethics, 16*(1), 39–66.

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology, 38*(1), 129–166.

Guglielmo, S., & Malle, B. F. (2019). Asymmetric morality: Blame is more differentiated and more extreme than praise. *PLoS One, 14*(3), Article e0213544.

James, A. (2012). Contractualism's (not so) slippery slope. *Legal Theory, 18*(3), 263–292.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263–291.

Kogut, T., & Ritov, I. (2005). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making, 18*(3), 157–167.

Kumar, R. (2015). Risking and wronging. *Philosophy and Public Affairs, 43*(1), 27–51.

Lewis, J., & Simmons, J. P. (2020). Prospective outcome bias: Incurring (unnecessary) costs to achieve outcomes that are already likely. *Journal of Experimental Psychology: General, 149*(5), 870–888.

Monin, B., & Merritt, A. (2012). *Moral hypocrisy, moral inconsistency, and the struggle for moral integrity*.

Mullen, E., & Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology, 67*.

Müller-Trede, J., Sher, S., & McKenzie, C. R. M. (2018). When payoffs look like probabilities: Separating form and content in risky choice. *Journal of Experimental Psychology: General, 147*(5), 662–670.

Oberdiek, J. (2017). *Imposing risk: A normative framework*. Oxford: Oxford University Press.

Pachur, T., & Kellen, D. (2013). Modeling gain-loss asymmetries in risky choice: The critical role of probability weighting. *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*, 3205–3210.

Peeters, G. (1971). The positive-negative asymmetry: On cognitive consistency and positivity bias. *European Journal of Social Psychology, 1*(4), 455–474.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.

Ryazanov, A. A., Knutzen, J., Rickless, S. C., Christenfeld, N. J., & Nelkin, D. K. (2018). Intuitive probabilities and the limitation of moral imagination. *Cognitive Science, 42*(S1), 38–68.

Scanlon, T. M. (1998). *What we owe to each other*. Cambridge, MA: Harvard University Press.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion, 16*(6), 862.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.

Shou, Y., & Song, F. (2017). Decisions in moral dilemmas: The influence of subjective beliefs in outcome probabilities. *Judgment and Decision making, 12*(5), 481–490.

Slovic, P. (2007). *Psychic numbing and genocide*. Psychological Science Agenda. , November https://www.apa.org/science/about/psa/2007/11/slovic.html.

Slovic, P. (2010). If I look at the mass I will never act: Psychic numbing and genocide. In *Emotions and risky technologies* (pp. 37–59). Dordrecht: Springer.

Thomson, J. J. (1990). *The realm of rights*. Cambridge, MA: Harvard University Press.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty, 5*, 297–323.

Västfjäll, D., Slovic, P., Mayorga, M., & Peters, E. (2014). Compassion fade: Affect and charity are greatest for a single child in need. *PLoS One, 9*(6).