Journal of Experimental Social Psychology 108 (2023) 104499

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Beyond killing one to save five: Sensitivity to ratio and probability in moral judgment☆

Arseny A. Ryazanov, Shawn Tinghao Wang, Dana Kay Nelkin *, Craig R.M. McKenzie, Samuel C. Rickless

University of California San Diego, United States of America

ARTICLE INFO

Keywords:
Morality
Probability
Risk
Decision-making
moral dilemma

ABSTRACT

A great deal of current research on moral judgments centers on moral dilemmas concerning tradeoffs between one and five lives. Whether one considers killing one innocent person to save five others to be morally required or impermissible has been taken to determine whether one is appealing to consequentialist or non-consequentialist reasoning. But this focus on tradeoffs between one and five may obscure more nuanced commitments involved in moral decision-making that are revealed when the numbers and ratio of lives to be traded off are varied, and when the probabilities of each outcome occurring are less than certain. Four studies examine participants' reactions to scenarios that diverge in these ways from the standard ones. Study 1 examines the extent to which people are sensitive to the ratio of lives saved to lives ended by a particular action. Study 2 verifies that the ratio rather than the difference between the two values is operative. Study 3 examines whether participants treat probabilistic harm to some as equivalent to certainly harming fewer, holding expected ratio constant. Study 4 explores an analogous issue regarding the sensitivity of probabilistic saving. Participants are remarkably sensitive to expected ratio for probabilistic harms while deviating from expected value for probabilistic saving. Collectively, the studies provide evidence that people's moral judgments are consistent with the principle of threshold deontology.

Normative ethics studies the principles of morally permissible and morally forbidden conduct, and typically tests those principles against people's reactions to particular (real or hypothetical) scenarios in order to achieve reflective equilibrium, or, in other words, a coherent and well-justified set of beliefs that includes general moral principles (Rawls, 1971; Thomson, 1990; Kamm, 1996). According to the method of reflective equilibrium, it matters greatly to moral theory which moral judgments human beings actually make in particular cases. This is not because "is" implies "ought"; the fact that most people judge that a particular course of action is morally permissible is not sufficient, on its own, to establish that that course of action is morally permissible. But reflective equilibrium treats moral judgments about cases as important sources of evidence for moral principles, to be weighed against, among other things, how intuitive those moral principles are independently of what those principles entail, and how well the moral principles, so justified, work together. A great deal of previous research has focused on whether or under what conditions people reason in ways consistent with consequentialist moral theory or not. In this article, we present data from five experiments to argue that threshold deontology – an under-explored non-consequentialist moral theory – provides a promising and unifying framework for capturing a wide set of intuitive moral judgments.

1. Moral judgment: Consequentialism vs. non-consequentialism

Two major normative ethical theories dominate the field: consequentialism and non-consequentialism. Consequentialists claim, roughly, that an act or omission is permissible (or required) if and only if its performance would lead to optimal results, i.e., consequences that are, on balance, at least as good as the consequences of any available alternative course of conduct (e.g., Sinnott-Armstrong, 2015). In deciding which action to adopt, the consequentialist looks to the total value of the outcome of each action, a result of weighing the harms and benefits. Non-consequentialists simply deny consequentialism. Most

☆ This paper has been recommended for acceptance by Dr. Paul Conway.

* Corresponding author at: Department of Philosophy, University of California, San Diego, La Jolla, CA 92093-0119, United States of America.
E-mail address: dnelkin@ucsd.edu (D.K. Nelkin).

https://doi.org/10.1016/j.jesp.2023.104499
Received 11 July 2022; Received in revised form 1 June 2023; Accepted 2 June 2023
Available online 24 June 2023
0022-1031/© 2023 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

non-consequentialists do not think that the value of consequences is morally irrelevant, but insist that, in addition to consequentialist considerations, other principles play an important role in ethical theory, and that *how* an outcome is achieved can make a moral difference. For example, some non-consequentialists appeal to principles such as the doctrine of doing and allowing (DDA: roughly, the view that it is more difficult to justify doing harm than it is to justify merely allowing harm) or the doctrine of double effect (DDE: roughly, the view that it is more difficult to justify intending harm than it is to justify merely foreseeing harm) (e.g., Nelkin & Rickless, 2014; Quinn, 1989a; Quinn, 1989b; Rickless, 1997).

For many years, normative ethical theories have been tested by means of thought-experiments that involve a forced choice between alternatives that lead to results of differing value. Consequentialists are well-positioned to appeal to rescue cases, in which one is forced to choose between rescuing one person and rescuing five (otherwise similar) people, claiming that it is at least permissible, and even required, to save the five rather than the one. Such cases seem to support the idea that one ought to act in a way that produces the best outcome (Kagan, 1989). But the question arises whether this principle generalizes and applies beyond these sorts of cases. In suggesting that it does not, non-consequentialists have appealed to other cases with the same balance of gain (typically, five lives) and loss (typically one life), but in reaction to which most judge that saving the five is morally impermissible. Examples include driving over one person trapped on the road ahead in order to save five people who are drowning in a lake at the end of the road (Quinn, 1989a), pushing a large man off a bridge above a train track in order to use his body to stop an oncoming train from crushing five people who are trapped on the track just beyond the bridge (Foot, 1967; Thomson, 1976; Thomson, 2008; Fitzpatrick, 2009), or chopping up a healthy patient in order to transplant his organs into five patients who will die without the appropriate organ transplant (Foot, 1967; Thomson, 1976).

2. Non-consequentialism: Absolutist vs. threshold deontology

The latter cases, in which there is a doing or intending of harm to one person in order to save five people, have led some non-consequentialists to what might be called "absolutist deontology," the view that non-consequentialist principles make it morally impermissible to engage in *any* conduct that harms, or that involves the intention to harm, *any* number of people in order to save a larger number from the same kind of harm. However, under pressure to accommodate cases in which the alternative to doing or intending harm to a small number of people would be catastrophic, many non-consequentialists have suggested that it is morally permissible to do or intend harm if one's conduct leads to an amount of good beyond a specified threshold. This alternative has come to be known as "threshold deontology" (Moore, 1997). Thus, whereas an absolutist deontologist would say that it is morally impermissible to kill one person even if that is required to save the people of a large metropolis, a threshold deontologist would say that killing the one is permissible if the amount of good that would result from the killing lies above a particular threshold (Alexander & Moore, 2016).

To this point, there has been a great deal of empirical research on how people respond to classic moral dilemmas. The idea that threshold deontology could be an implicit moral theory has received relatively little attention, and we aim to explore whether more attention to this possibility can help us to explain phenomena that are otherwise difficult to explain, and how it gives rise to new questions which can themselves be studied systematically. Because of the way the debate is often framed, much of the current empirical literature either ignores the possibility of a principled implicit moral theory of threshold deontology or assumes a set of options that rules it out from the start.

3. Dual process theory and its limitations

Studies in the existing literature typically employ hypothetical cases in which it is impossible to save five without killing (or intending to kill) one, often with an eye to informing the debate between consequentialists and non-consequentialists. Recent studies on dilemmas of this kind have suggested that people's moral judgments can systematically vary according to a variety of factors, such as the existence of physical contact (Cushman, Young, & Hauser, 2006), the intentional structure of the action (Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006), individual differences on working-memory-capacity tasks (Moore, Clark, & Kane, 2008), personality traits (Arvan, 2013), and the value one places on the agent (De Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017).

But at a more general level, empirical researchers have proposed several dual-process theories of moral cognition, according to which judgments in moral dilemmas are based on two competing processes: an outcome-based (or model-based) process responsible for consequentialist judgments, and an action-based (or model-free) process responsible for non-consequentialist judgments (e.g., Crockett, 2013; Cushman, 2013; Cushman et al., 2006; Cushman & Greene, 2011; Cushman, Young, & Greene, 2010; Greene et al., 2009; Greene & Haidt, 2002; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001; Paxton, Ungar, & Greene, 2012).

Many studies, however, assume that dual processes must reflect two distinct, competing moral principles, consequentialism and absolutist deontology, ignoring the possibility of a principled threshold deontology at work. This oversimplification can also be seen in researchers' tendency to categorize moral judgments as either "consequentialist" or "non-consequentialist" (see Kahane, Everett, Earp, Farias, & Savulescu, 2015, who have challenged this tendency, but based on reasons different from the ones we provide, as well as Conway, Goldstein-Greenwood, Polacek, & Greene, 2018; Conway & Gawronski, 2013).

What this dichotomy obscures is that those opposed to killing at a particular difference or ratio of good done to harm done may in a principled way shift to endorsing killing at a higher such difference or ratio. For example, of those who reject killing one to save five, some may endorse killing one to save ten. And of those who accept killing one to save five, some may reject it when the number killed grows to four. Such shifts would suggest a more nuanced principle of morality, such as threshold deontology, that is consistent with the integration of outcome-based and action-based processes (Cohen & Ahn, 2016; Hutcherson, Montaser-Kouhsari, Woodward, & Rangel, 2015). Threshold deontology is in tension with the simple kind of dual process account described above insofar as it seems antecedently unlikely that an emotional system will perfectly track only those cases in which the numbers of those killed and those saved are both relatively small. The idea that threshold deontology might play an explanatory role in moral decision-making also points to the possibility of rational consistency that is not otherwise captured by a simple dual process account.

4. Threshold deontology vs. dual process: Getting away from extreme cases

One reason that threshold deontology might be ignored is that it is often associated only with extreme cases. Philosophers who are threshold deontologists often appeal to cases in which the only available alternative is catastrophic (Nozick, 1974; Fried, 1978; Nagel, 1979; Moore, 1997; though see Thomson, 1990 and Brennan, 1995). An example would be the killing of one person in order to avoid the destruction of a large city or an entire nation.

Interesting psychological research mirrors this focus on extreme cases: notable deviations from the paradigmatic 1 vs 5 scenarios appear

in psychological research where participants are asked to evaluate the killing of a person in order to avoid a catastrophe (Bartels, 2008; Nichols & Mallon, 2006). This work leaves open the possibility that the threshold for overriding genuine deontological constraints is much lower than at catastrophic levels. In the studies described below, we explore the power of threshold deontology to explain participants' responses to a variety of dilemmas, including ones that involve differences and ratios that are both lower than in the classic 1 vs 5 cases, but also higher without being so high as to compare one life to that of the population of a city or nation.

While recent research suggests that moral judgments are sensitive to a total weighing of harms and benefits, it remains an open question just how they are sensitive to it in cases where one must cause harm to achieve a greater benefit as in classic moral dilemmas. For example, Tassy, Oullier, Mancini, and Wicker (2013), as well as Shenhav and Greene (2010), found that participants were sensitive to the number of people who could be saved in dilemmas where participants could choose one of two groups to save. However, neither of these tasks constituted sacrificial moral dilemmas, in that they involved choosing to benefit one party at the expense of another (picking between positive outcomes), rather than imposing harm on one party for the benefit of another. Trémolière and Bonnefon (2014) found that more participants made "utilitarian" judgments when the number of people who could be saved by acting increased. However, some of the scenarios used did not present true sacrificial moral dilemmas. Rather, in their scenarios, failure to act would result in everyone dying, including the one who would have been killed (slightly earlier) to save others. Costa-Lopes, Mata, and Mendonça (2021) found that participants treated cases with different numbers of potential victims differently when identifying information was provided. However, the scenarios employed were also not classic sacrificial moral dilemmas that required killing one to save others in the robust sense, but rather required the shifting or diverting of a causal sequence already in place. (See Foot (1984), Thomson (1985), and Rickless (1997) for discussion of this distinction.)

The current Study 1, by presenting participants with several different sacrificial dilemmas with different ratios of harm to benefit, contributes toward filling in this lacuna, and examines whether participants behave in accord with threshold deontology at non-catastrophic levels.

5. Fixing the threshold: Difference vs. ratio under certainty

Even if participants exhibit patterns of response that are consistent with threshold deontology, two questions arise about the factors that underlie moral decisions. First, how is the appropriate threshold determined for any given course of conduct? Although many options are logically possible, the main alternatives in the relevant debate are two: the threshold might look to the *difference* between, or to the *ratio* of, the amount of good and the amount of harm to be achieved by the relevant conduct. Thus, in the case of killing a certain number of people ($N_{\text{lives ended}}$) to save a certain number of people ($N_{\text{lives saved}}$) the *difference* threshold deontologist would say that the killing is morally permissible if and only if $N_{\text{lives saved}} - N_{\text{lives ended}}$ is above a certain number, whereas the *ratio* threshold deontologist would say that the killing is morally permissible if and only if $N_{\text{lives saved}} / N_{\text{lives ended}}$ is above a certain number. Thus, we define two distinct formulas that could be operative, where R represents ratio and D represents difference:

$$R = N_{\text{lives saved}} / N_{\text{lives ended}}$$

$$D = N_{\text{lives saved}} - N_{\text{lives ended}}$$

Study 2 presents participants with scenarios designed to test whether their responses are more consistent with R or with D.

6. Fixing the Threshold: Difference vs. Ratio Under Uncertainty

A second question for threshold deontology, and for non-consequentialists more generally, is how risk and uncertainty should factor in moral decision-making. With respect to the influence of uncertainty on moral judgment, most moral dilemma studies have focused on actions whose outcomes are described as certain to happen (or at least never described as uncertain). In real life, however, we rarely know with certainty what will happen if we act one way rather than another, and often work with probabilities somewhere between 0 and 1. Moreover, in psychological research, even when participants are told that outcomes are certain, there is evidence that they often substitute their own probability estimates of less than 100% for outcomes that are described as certain (Ryazanov, Knutzen, Rickless, Christenfeld, & Nelkin, 2018; Shou & Song, 2017).

Consequentialists have a simple answer to how we should act when we confront uncertain outcomes, since decisions should depend entirely on a weighing of harms and benefits. Such a weighing can be easily understood in terms of a calculation of expected value (or, in other words, the result of multiplying the probability and value of each possible outcome). In fact, the most influential forms of consequentialism are often presented as the view that one should perform the action with the highest expected value.

The situation is not nearly as clear in the case of non-consequentialism in the form of either absolute or threshold deontology, where the value of outcomes is not the entire determinant of what one ought to do. For example, it is not obvious from the perspective of threshold deontology what are the contours of moral permissibility when the probability of harm (or benefit) is low but the harm (or benefit) is very significant, or when the probability of harm (or benefit) is high but the harm (or benefit) is much less significant. Some non-consequentialists have attempted to address the issue of risk and uncertainty (see Hansson, 2003; Oberdiek, 2017), but the issue remains a live one. What is clear, however, is that for both the consequentialist and the threshold deontologist, outcomes play *some* role in our moral decision-making. Putting together the question of ratio vs. difference in the context of uncertainty of outcomes leads us to define expected ratio and expected difference.

We define the expected ratio (ER) here in a way that incorporates the number of people who might be saved and the probability (P) that they will be saved, as well as the number of people at risk of being killed and the probability that they will be killed. We thus first need to define the expected value of lives saved ($EV_{\text{lives saved}}$), which is a product of the lives that might be saved and the probability that they are, and the expected value of lives ended ($EV_{\text{lives ended}}$), which is a product of the lives that might be ended and the probability that they are:

$$EV_{\text{lives saved}} = N_{\text{lives saved}} \text{ x } P_{\text{lives saved}}$$

$$EV_{\text{lives ended}} = N_{\text{lives ended}} \text{ x } P_{\text{lives ended}}$$

With these variables defined, we can now give a complete definition of expected ratio:

$$ER = EV_{\text{lives saved}} / EV_{\text{lives ended}}$$
$$= (N_{\text{lives saved}} \times P_{\text{lives saved}}) / (N_{\text{lives ended}} \times P_{\text{lives ended}})$$

We define the expected difference (ED) as follows:

$$ED = EV_{\text{lives saved}} - EV_{\text{lives ended}}$$
$$= (N_{\text{lives saved}} \times P_{\text{lives saved}}) - (N_{\text{lives ended}} \times P_{\text{lives ended}})$$

Mikhail (2011) hypothesized that participants' moral grammar includes a "moral calculus of risk." But Mikhail (2011) did not test whether the moral calculus of risk governs participants' judgments. In our studies, we test whether the expected difference between, or the expected ratio of, good and bad outcomes plays a role in moral judgment.

Researchers have begun to study the role of probability in moral reasoning (Ryazanov, Wang, Rickless, McKenzie and Nelkin, 2021). Fleischhut, Meder, and Gigerenzer (2017) found that moral judgments when outcomes are certain to occur differ from when those outcomes are uncertain, though without specifying any probability for the outcomes' occurrence. In addition to varying the number of lives that could be saved, Shenhav and Greene (2010) varied the probability that the latter group of people do not actually need saving (e.g., the probability that a group of people blocked in an office building will successfully escape anyway). There are two additional probabilities that are relevant, but were not examined: the probability that the plan of saving them will be successful, and the probability that the plan will kill a number of people. Do these probabilities matter? How will they interact with the role of expected value? These questions remain unanswered, and are important for testing whether there is a moral difference between harming and omitting to save, as is presupposed by the non-consequentialist principle, the DDA.

7. Moral judgment under uncertainty: Risk seeking vs. risk aversion

Though moral dilemmas differ from self-interested dilemmas, people in the latter cases often do not behave in accord with expected value, and the same might be true for the former cases. In particular, people tend to be risk seeking when dealing with losses (they often prefer a gamble to a sure loss with the same expected value) and they tend to be risk averse when dealing with gains (they often prefer a sure gain to a gamble with the same expected value). For example, in the "Asian disease problem," participants must decide between certain losses of life and probabilistic losses of life, as well as between certainly saving a group of individuals and probabilistically saving a group of individuals (Tversky & Kahneman, 1981). For the loss of life scenario participants prefer the risky option, but for the saving lives scenario participants prefer the certain option. These non-sacrificial dilemmas suggest that probabilities and expected value may matter in sacrificial moral dilemmas, and that, furthermore, probabilistic harm and saving may be treated differently (Diederich, Wyszynski, & Ritov, 2018).

There has been a growing consensus that moral judgments depend on domain-general principles such as those involving causal and intentional attribution (Cushman & Young, 2011), language (Costa et al., 2014), psychological essentialism (De Freitas, Cikara, Grossmann, & Schlegel, 2017, 2018), and efficiency (De Freitas & Johnson, 2018). Thus, it is promising that risk, as a domain-general factor, would influence people's judgments in moral dilemmas as well. If people are risk seeking for losses, they might be willing to accept a greater expected number of people being killed when the harm is probabilistic rather than certain. And if they are risk averse for gains, they might need a greater expected number of people being saved when the saving of lives is probabilistic rather than certain. It remains to be seen whether sacrificial moral dilemmas treat harm as a loss, and benefit as a gain, and, if so, whether participants are risk seeking for probabilistic harm and risk averse for probabilistic saving.

So, in addition to departing from 1 vs 5 cases and from the focus on catastrophic alternatives, we also depart from the presumption of certainty to test whether participants treat equivalent expected values similarly when probabilities of harm or rescue differ. In study 3, we examine whether participants treat probabilistic harm to some as equivalent to certainly harming fewer, when expected values are held constant; in study 4, we focus on probabilistic saving rather than harm.

8. Overall aims

Thus, this paper has two main aims: (1) to test the hypothesis that participants exhibit judgments consistent with threshold deontology, rather than consequentialism or absolutist deontology; and (2) to systematically examine the role of expected value and probability in moral judgment, which further involves testing (i) the relative role of ratio and difference of number when it comes to trading off harm and benefits in moral judgments, (ii) how the expected number of people being saved and the expected number of people being killed interact with each other, and (iii) how varying the probability of harming and successfully saving affects participant responses.

9. Study 1 – Sensitivity to ratio in abstract moral dilemmas

Study 1 examines whether participants are sensitive to the ratio of lives saved to lives ended in an abstract moral dilemma, to determine whether moral judgments are sensitive to the ratio of harm to good done, which is consistent with threshold deontology, or whether participants either endorse or reject action according to principles insensitive to ratio (such as consequentialism or absolutist deontology). In addition, Study 1 begins to explore whether participants are sensitive to expected ratio vs. expected difference between lives saved and lives ended, and whether participants are risk seeking when harm is probabilistic rather than certain.

9.1. Study 1 materials and participants

Participants read one of four scenarios, in which a group of people could be saved from a terrorist attack by harming a bystander. As an example, one of the scenarios was as follows:

> *Harry has a plan to save two people from a terrorist attack. For Harry's plan to work it is necessary that Robert, a bystander, is killed. If Harry carries out this plan, he will certainly save the two people, but in doing so will certainly kill the bystander.*

Participants saw a version of the scenario where one would be killed to save two (expected ratio (ER) = 2), one would be killed to save one hundred (ER = 100), one would have a 1 in 50 (2%) chance of being killed to save two (ER = 100), or 99 would be killed to save 100 (ER ≈ 1.01). 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 will certainly save two people, but will certainly kill a bystander?* (−5: *very confident he should not carry out the plan*, to 5: *very confident he should carry out the plan*). We chose this dependent variable, rather than a continuous forbidden-obligatory measure, because according to many consequentialists and non-consequentialists alike, actions cannot be *more or less* morally right or wrong (see Ryazanov et al., 2018), and questions that don't explicitly ask about confidence (e.g., agree-disagree) implicitly do by asking for degree of agreement/disagreement. We used a single scale measure, rather than process dissociation as endorsed by Conway and Gawronski (2013), because we are interested in whether judgments change at different expected values, rather than in the strength of deontological and consequentialist "inclinations" that could contribute to moral judgments. One hundred and twenty-two participants were recruited via Amazon's Mechanical Turk (110 passed an attention check and were retained for analysis (90.1% passed); mean age = 32.9, SD = 8.30; 56.4% of our sample identified as female, 43.6% as male). We report all measures, manipulations, and exclusions in this and all subsequent studies other than additional demographic measures. The attention check involved a statement embedded in a second scenario that prompted participants to select a specific number on scale instead of responding to the question prompt. Sample size was determined prior to data collection, and was selected to be able to detect a medium-large effect size for the effect of expected ratio, which was determined to require >87 participants ($f = 0.30$, alpha = 0.05, beta = 80%, two-tailed) using Gpower software (Faul, Erdfelder, Lang, & Buchner, 2007). Sensitivity power analyses for Study 1 (using alpha = 0.05 and beta = 80%, two-tailed) determined a minimum detectable effect size (MDES) of $f = 0.27$, using Gpower software (Faul et al., 2007).

9.2. Study 1 Results

We began by examining how responses to whether the action should be carried out correspond to the action's expected ratio. A one-way ANOVA yielded a significant effect of condition on moral judgment, $F(3,106) = 15.8$, $p < .001$, $\eta^2 = 0.310$ (mean ER 1.01 [kill 99 to save 100] $= -2.23$, SD $= 2.67$; mean ER 2 [kill 1 to save 2] $= 0.111$, SD $= 2.45$; mean ER 100 [kill 1 to save 100] $= 2.20$, SD $= 2.93$; mean ER 100 [2% chance kill 1 to save 2] $= 2.04$, SD $= 2.70$); see Fig. 1). Planned contrasts found ratings with ER 1.01 differed significantly from those with ER 2 (Welch $t(50.2) = -3.32$, Holm-adjusted $p = .005$, $d = 1.30$) and ratings with ER 2 differed significantly from ER 100 [kill 1 to save 100] (Welch $t(54.7) = -2.93$, Holm-adjusted $p = .010$, $d = 0.37$), while ratings with ER 100 [killing 1 to save 100] did not differ significantly from ER 100 [2% chance of killing 1 to save 2] (Welch $t(54.97) = 0.219$, Holm-adjusted $p = .828$, $d = 0.31$). Thus, participants were sensitive to the expected ratios presented to them, but not to the same expected ratio presented through a different probability: kill one to save 100 and 2% chance of killing one to save two.

We also categorized responses into "should act" (responses >0) and "should not act" (responses ≤ 0) to examine whether the observed sensitivity to expected ratio is limited to confidence in action, or extends to binary decisions regarding whether to act or not. A logistic regression revealed an effect of condition on binary decisions, $\chi^2(3, N = 110) = 24.5$, $p < .001$ (proportions voting in favor of action: 0.269, 95% CI [0.136,0.464] when ER $= 1.01$; 0.481, 95% CI [0.308,0.660] when ER $= 2$; 0.800, 95% CI [0.622,0.907] when ER 100 was expressed as kill 1 to save 100; 0.815, 95% CI [0.627,0.921] when ER 100 was expressed as a 2% chance of killing 1 to save 2). Consistent with linear responses, planned contrasts revealed that the proportion favoring action with ER $= 1.01$ did not differ significantly from when ER $= 2$ (OR $= 0.397$, 95% CI [0.126,1.252]), but that ER $= 2$ did differ significantly from when ER $= 100$ [kill 1 to save 100] (0.232 [0.072,0.748]), and that ER 100 [kill 1 to save 100] did not differ significantly from those with ER 100 [2% chance of killing 1 to save 2]. Thus, for binary decisions, we observed a similar pattern to the results observed for continuous decisions.

9.3. Study 1 Discussion

Study 1 found evidence for sensitivity to expected ratio. Participants more often endorsed an action that harmed one to save others when the ratio regarding the expected value of lives saved to the expected value of lives ended was larger. This sensitivity suggests that people are neither absolutist deontologists nor consequentialists, instead making decisions consistent with the principle of threshold deontology. Notably, it is not only in cases of catastrophic harm that participants' responses are consistent with the overriding of deontological constraints against harming. Rather, we see significant shifts in mean level of confidence from a negative valence in the case of killing 99 to save 100 to a neutral mean level of confidence in the case of killing one to save two, to a mean level of positive valence in the case of killing one to save 100. This shows that when keeping all else equal, the numbers are important in non-catastrophic cases in shifting judgments. This suggests that to the extent that people do take there to be thresholds at which the magnitude of harm can override deontological constraints, these points are reached in far more cases than is suggested by typical illustrations of threshold deontology.

Study 1 also includes preliminary evidence regarding sensitivity to probabilistic forms of the same expected ratio. Participants treated the two scenarios whose expected ratios were identical no differently, despite one of them involving a probabilistic harm and one a certain harm. In this instance at least, moral uncertainty didn't have any impact independent of the expected ratio on participants' application of their ethical principles. However, we should be cautious in interpreting this finding, given the small sample size used to explore sensitivity to probability—Study 3 explores this finding with a larger sample capable of detecting smaller effects.

But first, we note that the results of the first study suggest that it is expected ratio rather than expected difference that matters when it comes to harming some to save others. It is possible that such decisions could be made not on the ratio, but instead on the number of lives gained. That is, killing 99 to save 100 involves a net gain (or difference) of one life, and in this way is similar to killing one to save two. However, our data indicate that the latter option is regarded much more favorably

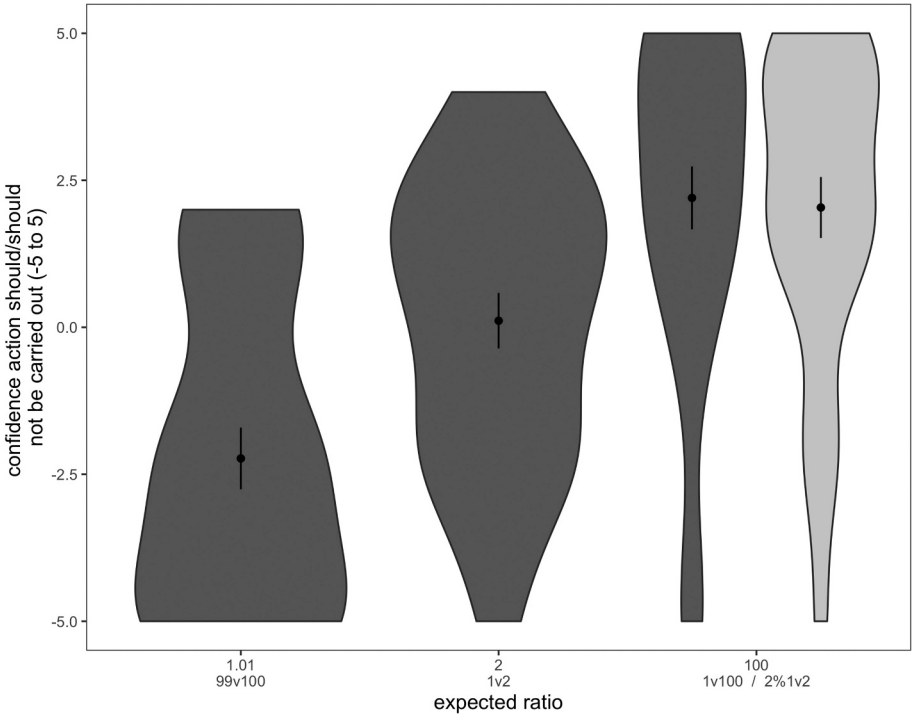

Fig. 1. Sensitivity to expected ratio of lives saved to lives lost in ratings of confidence in action. Error bars represent one standard error.

than the former. In Study 2, we set out to test more directly the hypothesis that it is in fact ratio rather than difference that is operative.

10. Study 2a – Insensitivity to difference in value of abstract moral dilemmas

Study 2a sought to verify that participants are insensitive to difference between the number of lives saved and lost, by manipulating ED while holding ER constant.

10.1. Study 2a Materials and participants

Each participant read about one of three plans, which were identical to those of Study 1 except for the numbers involved. The plans involved killing 1 to save 2 (difference = +1 life), killing 10 to save 20 (difference = +10 lives), or killing 100 to save 200 (difference = +100 lives). Thus, ED is manipulated, while ER is constant at 2. One hundred and fifty-eight participants were recruited via Amazon's Mechanical Turk (135 passed an attention check and were retained for analysis (85.4% passed); mean age = 34.2, SD = 10.8; 63.7% of the sample identified as female; 36.3% as male). Sample size was determined prior to data collection, and was selected to be able to detect a medium-large effect size for the effect of expected difference, which was determined to require >87 participants (f = 0.30, alpha = 0.05, beta = 80%, two-tailed) using Gpower software (Faul et al., 2007). Sensitivity power analyses for Study 2a (using alpha = 0.05 and beta = 80%, two-tailed) determined a minimum detectable effect size (MDES) of f = 0.24 using Gpower software (Faul et al., 2007).

10.2. Study 2a Results

We examined differences in responses to each of the three scenarios. If participants favor the action more when it saves more net lives, there should be a sharp increase in supporting the action as the net number goes from +1 to +10 to +100 lives. If, instead, participants are sensitive to the expected ratio, then the three scenarios should be regarded as effectively identical, with the action in each one saving twice as many as are sacrificed. A one-way ANOVA confirmed that there was no significant difference between the scenarios, $F(2,132) = 0.268$, $p = .765$, $\eta^2 = 0.00405$ (Mean 1v2 = −0.156, SD = 3.10; Mean 10v20 = 0.00, SD =

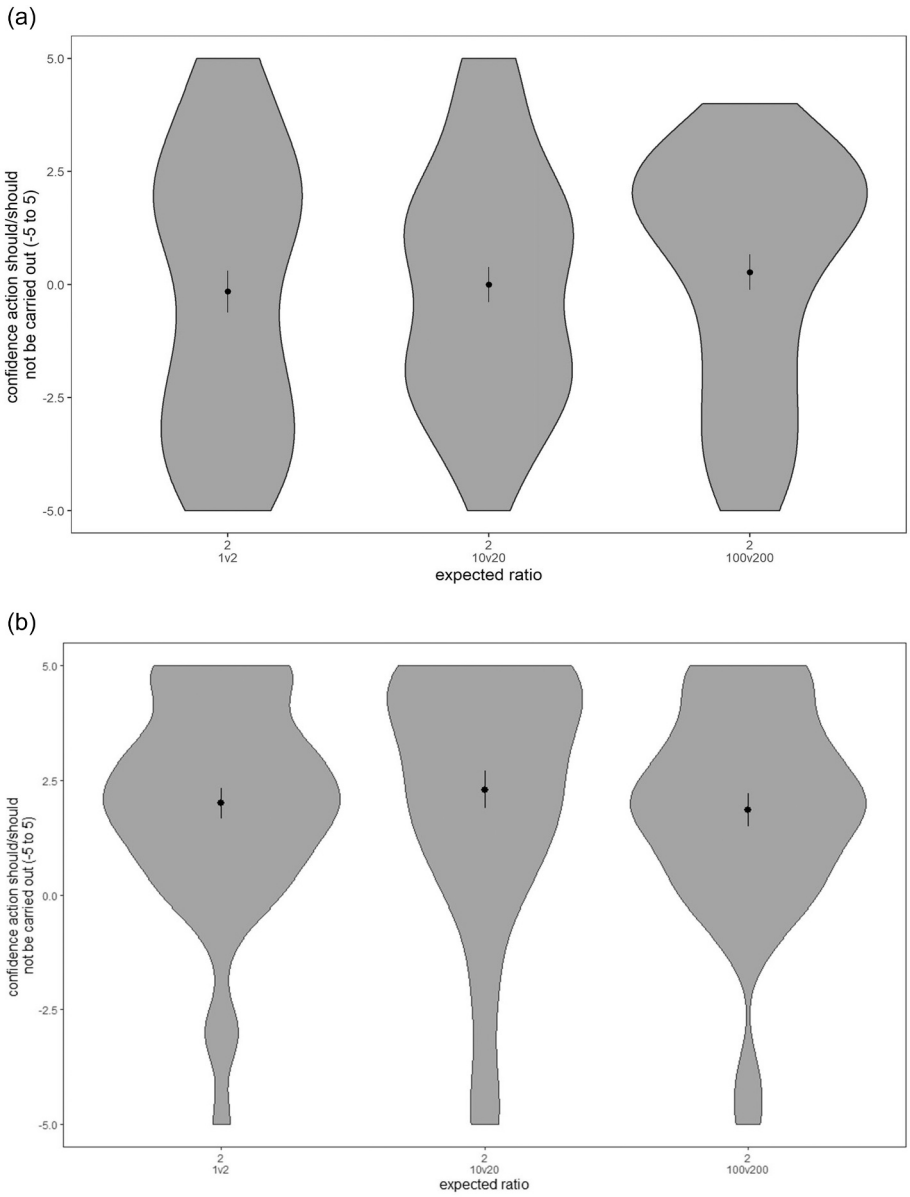

Fig. 2. a. Insensitivity to expected difference between lives lost and saved, while holding expected ratio of 1 to 2 constant, in ratings of confidence in action in abstract scenarios. Error bars represent one standard error.
b. Insensitivity to expected difference between lives lost and saved, while holding expected ratio constant, in ratings of confidence in action in concrete scenarios. Error bars represent one standard error.

2.63; Mean 100v200 = 0.273, SD = 2.60; see Fig. 2a). Furthermore, a logistic regression with ED entered as categorical variable revealed no effect on binary decisions after categorizing decisions into "should act" and "should not act", $\chi^2(2, N = 135) = 1.39$, $p = .500$ (proportions in favor of acting:

0.591, 95%CI [0.444, 0.723] when ED = 100; 0.478 95% CI [0.342, 0.618] when ED = 10; 0.489 [0.350, 0.630] when ED = 1).

10.3. Study 2a Discussion

The lack of sensitivity to the expected difference between the number of lives lost and the number of lives saved suggests that participants in Study 1 were sensitive to expected ratio, instead of the net gain in lives from the action (or, even the raw number killed or saved). Given that moral theorists putting forward threshold deontology as the normatively correct theory have not offered precise proposals about how thresholds should be determined, it is not possible to compare this result to any extant well-worked-out moral theory. However, it might be argued that consistency in moral reasoning should favor ratio over difference as the determinant of the threshold. The reason is that one might see the trade-off between 100 lives ended and 200 lives saved as a series of pairwise tradeoffs between one and two. This is an issue worthy of further study by moral theorists, and the results here can inform it.

The scenario we used was a fairly abstract one, which, while specifying the numbers involved, did not flesh out how the fewer would die, nor how that would save the many. It is possible that the use of abstract scenarios encouraged a certain kind of calculation that would not be elicited by scenarios with more detail, as would be consistent with some recent work on Construal Level Theory and moral judgment (Gong & Medin, 2012; Lammers, 2012), and the finding that participants are less willing to act in abstract situations than concrete ones (Agerström & Björklund, 2009; Amit & Greene, 2012). Other work has suggested a more complex picture in which construal level interacts with other factors such as time and cognitive load (Körner & Volk, 2014). While intriguing, all but one of the scenarios used in this work were not sacrificial dilemmas in which one person, who would otherwise live, could be killed to save others. In these scenarios, instead, the person who would be killed was either fatally injured or would be killed with the others in the absence of action. So it is unclear how the relevant factors would interact in true sacrificial dilemmas. However, even if multiple factors in addition to construal level affect moral judgments in interacting ways, it is important to learn whether the fact that the scenarios are abstract in Study 2a made a distinctive contribution to the particular results. Thus, we sought to replicate our effects also with more detailed scenarios.

11. Study 2b – Insensitivity to difference in expected value of concrete moral dilemmas

Study 2b sought to extend the finding in Study 2a that participants are insensitive to the expected difference between benefit and harm in scenarios that have concrete details.

11.1. Study 2b Materials and participants

Concrete scenarios were created in which the expected ratio of lives saved to lives lost was held constant, though the expected difference in numbers between the two groups varied, to examine whether it is the ratio rather than the difference of lives saved to lives lost that is operative. Each participant again read about one of three plans: killing 1 to save 2, killing 10 to save 20, or killing 100 to save 200, though now, instead of more abstract plans, we utilized a more detailed scenario that involved setting off an explosion to prevent a rocket from reaching a house. Subjects in the condition pitting sacrificing ten against saving twenty read the following scenario:

A missile has been accidentally fired at a house with 20 people in it. Bob is in charge of a missile defense tool that can destroy this missile by firing a rocket that can automatically detect the missile's location. The rocket will incapacitate the missile by setting off an explosion in the air near it. As Bob knows, the rocket's explosion near the missile will disable the missile, but will also kill 10 people standing in a field over which the missile will be intercepted. Firing the rocket given the timing and flight path of the missile is the only available option to prevent the missile from continuing on its path to the house with 20 people in it. Bob also knows the following facts. If Bob does not intervene, then the missile will certainly hit the house and kill all 20 people in it; if Bob intervenes, then the rocket Bob can set off will certainly destroy the missile and spare the people in the house, but will certainly kill the 10 people in the field.

Participants were asked, *Should Bob set off a rocket that he knows will kill 10 people, but that he also knows will destroy a missile that will otherwise kill 20 people?* (−5: *very confident he should not set off the explosion,* to 5: *very confident he should set off the explosion*). One hundred and forty-nine participants were recruited via Amazon's Mechanical Turk (129 passed an attention check and were retained for analysis (86.6% passed); mean age = 34.2, SD = 9.78; 54.7% of the sample identified as female; 45.3% as male). Sample size was determined prior to data collection, and was selected to be able to detect a medium-large effect size for the effect of expected ratio, which was determined to require >87 participants ($f = 0.30$, alpha = 0.05, beta = 80%, two-tailed) using Gpower software (Faul et al., 2007). Sensitivity power analyses for Study 2b (using alpha = 0.05 and beta = 80%, two-tailed) determined a minimum detectable effect size (MDES) of $f = 0.25$, using Gpower software (Faul et al., 2007).

11.2. Study 2b Results

As in the case of abstract scenarios, a one-way ANOVA confirmed that there was no significant difference between any of the more concrete scenarios, in which expected ratio was held constant, but expected difference was varied, $F(2, 125) = 0.353$, $p = .689$, $\eta^2 = 0.00593$. (Mean 1v2 = 2.00, SD = 2.25; Mean 10v20 = 2.30, SD = 2.72; Mean 100v200 = 1.85, SD = 2.27; see Fig. 2b). Furthermore, a logistic regression with ED entered as a categorical variable revealed no effect on binary decisions after categorizing responses into "should act" and "should not act", $\chi^2(2, N = 128) = 0.134$, $p = .935$ (proportions in favor of acting: 0.769, 95% CI [0.614,0.875] when ED = 100; 0.795, 95% CI [0.652,0.890] when ED = 10; 0.800, 95% CI [0.659,0.892] when ED = 1).

11.3. Study 2b Discussion

Again, participants remained insensitive to expected difference in lives lost and saved. Together with Study 1 and Study 2a, this suggests that participants are sensitive to ratio, rather than to difference, in number of lives saved to lives ended or to some combination of the two. Consistent with studies that find participants to be more willing to act in concrete than in abstract situations (Agerström & Björklund, 2009; Amit & Greene, 2012), an exploratory analysis revealed that participants expressed greater confidence in action in the concrete scenarios of Study 2b than in the abstract scenarios of Study 2a, $t(262) = 6.30$, $p < .001$, $d = 0.78$. Insensitivity to expected difference was apparent in both kinds of scenarios.

It should be noted that Studies 2a and 2b were designed to reveal a medium-large effect of expected difference, so finding a non-significant effect in each case is not particularly informative on its own. However, any effect of expected difference is clearly much smaller than the effect of expected ratio found in Study 1.

Thus far, we have found participant responses to be consistent with threshold deontology, and with thresholds that are sensitive to expected ratio rather than to expected difference. Study 1 also began to explore the effect of making outcomes uncertain, finding that participants

treated the trade-off of ending one life for saving 100 in a similar way to the 50% probability of ending one life and saving 50. In Study 3, we continued to examine how such judgments incorporate probability, beginning with comparing probabilistic harm to certain harm.

12. Study 3 – Sensitivity to probability of harm

Study 3 examined sensitivity to various probabilities through which an amount of harm could be expressed. For example, might participants be more confident in carrying out an action that risks a 1% chance of killing 100 people, than an action that certainly kills one, to benefit others? As mentioned, research on non-sacrificial dilemmas has demonstrated that participants often prefer risky losses to certain losses, holding expected value constant. If people are sensitive to the level of risk of harm in moral dilemmas, this would indicate that observed participant thresholds are sensitive not just to the ratio of harm to good done, but also to the likelihood of that harm occurring.

12.1. Study 3 Materials and participants

Study 3 examined the effect of probability across three expected ratios: ER = 5, ER = 2, and ER = 1.25. Probabilities of the people being sacrificed were varied, with expected ratio remaining constant. Probabilities explored included 1%, 10%, 20%, 50%, 80% (only for 4v5), and 100%, *see* Table 1 *for combinations*. We again utilized the more detailed scenarios from Study 2b, which involved setting off an explosion to prevent a rocket from reaching a house. In the ER = 5 cases, for example, participants were asked, *Should Bob set off a rocket that he knows will have a X% chance of killing Y persons* [where XY = 1], *but that he also knows will destroy a missile that will otherwise kill 5 people?* (−5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). 706 participants were recruited (616 passed an attention check and were retained for analysis (87.2% passed); mean age = 36.5, SD = 12.0, 61.1% identified as female; 38.9% as male). Each subject rated only one scenario, and provided brief demographic information. Sample size was determined prior to data collection, and was selected to be able to detect a small-medium effect of probability, which was determined to require >550 participants ($f = 0.12$, alpha = 0.05, beta = 80%, two-tailed, two-tailed) using Gpower software (Faul et al., 2007). Sensitivity power analyses for Study 3 (using alpha = 0.05 and beta = 80%) determined a minimum detectable effect size (MDES) of $f = 0.11$, using Gpower software (Faul et al., 2007).

12.2. Study 3 Results

A two-way ANOVA with probability and expected ratio entered as categorical variables revealed a significant effect of expected ratio, $F(2, 600) = 21.45$, $p < .001$, $\eta_p^2 = 0.066$ (mean 4v5 = 1.41, SD = 2.75; mean 1v2 = 1.86, SD = 2.45; mean 1v5 = 2.97, SD = 2.13; see Fig. 3a). Planned contrasts showed statistically significant differences in ratings by expected ratio condition (ER = 5 versus ER = 1.25: Welch t(426.91) = 6.646, Holm adjusted $p < .001$, $d = 0.63$; ER = 5 versus ER = 2: Welch t(364.46) = 4.697, Holm adjusted p < .001, $d = 0.49$; ER = 2 versus ER = 1.25: Welch t(418.62) = 1.782, Holm adjusted $p = .075$, $d = 0.17$). However, the ANOVA revealed that there was not a significant effect of probability, $F(5, 600) = 1.11$, $p = .353$, $\eta_p^2 = 0.009$, (mean 1% = 2.18, SD = 2.49; mean 10% = 1.68, SD = 2.74; mean 20% = 2.09, SD = 2.82; mean 50% = 2.12, SD = 2.82; mean 80% = 2.00, SD =2.45; mean 100% = 2.00, SD = 2.50), nor was there an interaction between probability and expected ratio, $F(8, 600) = 0.887$, $p = .527$, $\eta_p^2 = 0.012$, see Fig. 3b and Table 2. This suggests that, again, participants were sensitive to expected ratio, regardless of the probability of harm, even when the uncertain harm covered the range down to a 1% chance of occurrence. A logistic regression with expected ratio and probability entered as categorical variables confirmed that "should act"/ "should not act" decisions were sensitive to expected ratio, $\chi^2(2, N = 616) = 18.5$, $p < .001$ (72% would act when ER = 1.25; 74% when ER = 2; 88% when ER = 5 (proportions in favor of acting: 0.882, 95% CI [0.826,0.921] when ER = 5; 0.743, 95% CI [0.676,0.801] when ER = 2; 0.724, 95% CI [0.665,0.777] when ER = 1.25). Planned contrasts revealed significant differences between all ERs except ER = 2 vs. ER = 1.25 (ER = 5 vs ER = 1.25 OR = 2.838 [1.68, 4.80]; ER =5 vs. ER = 2 OR = 2.57 [1.48, 4.48]; ER = 2 vs ER = 1.25 OR = 1.10 [0.716, 1.70]). Consistent with Study 1 findings, we did not observe sensitivity to probability when the harm was expressed probabilistically, $\chi^2(5, N = 616) = 3.52$, $p = .619$ (proportions in favor of acting: 0.826, 95% CI [0.75,0.88] when probability = 1; 0.74 95% CI [0.65,0.81] when probability = 10; 0.80, 95% CI [0.71, 0.86] when probability = 20, 0.78 95% CI [0.69,0.84] when probability = 50, and 0.740 95% CI [0.58,0.85] when probability = 100). Finally, there was no interaction between expected ratio and harm probability, $\chi^2(8, N = 616) = 5.68$, $p = .684$, see Table 2.

Table 1
Study 3 scenarios.

ER	Probability of harm	Scenarios
1.25	1%	1% chance of killing 400 people to save 5 others
1.25	10%	10% chance of killing 40 people to save 5 others
1.25	20%	20% chance of killing 20 people to save 5 others
1.25	50%	50% chance of killing 8 people to save 5 others
1.25	80%	80% chance of killing 5 people to save 5 others
1.25	100%	100% chance of killing 4 people to save 5 others
2	1%	1% chance of killing 100 people to save 2 others
2	10%	10% chance of killing 10 people to save 2 others
2	20%	20% chance of killing 5 people to save 2 others
2	50%	50% chance of killing 2 people to save 2 others
2	100%	100% chance of killing 1 person to save 2 others
5	1%	1% chance of killing 100 people to save 5 others
5	10%	10% chance of killing 10 people to save 5 others
5	20%	20% chance of killing 5 people to save 5 others
5	50%	50% chance of killing 2 people to save 5 others
5	100%	100% chance of killing 1 person to save 5 others

12.3. Study 3 Discussion

Study 3 showed that participants were sensitive to expected ratio even across a fairly subtle range. Even with increased power from a larger number of participants, we continued to see no clear relationship between probability of harm, when expected ratio is fixed, and moral judgment. Regardless of how the expected ratio was presented to participants (e.g., 1% chance of killing 100 to save 5, or 1 certainly being killed to save 5), participants remained equally sensitive to the value. As expected ratio increased, participants expressed greater confidence that harmful action should be carried out. This is just as predicted by threshold deontology.

So far, while we varied the number of people involved on both the harming side and the saving side of the dilemma, we explored the effect of probability only on the harming side. We next explored how participants treat ethical dilemmas where the saving is certain versus probabilistic.

13. Study 4 – Insensitivity to probability of saving

Study 4 turned to a different probability: probabilistic saving with certain harm. We adapted the scenarios from Study 3 to examine a parallel range of probabilities, this time on the saving side. For example, would an action that kills one to save two be judged differently from an action that kills one to save four who have a 50% chance of dying without the intervention? This uncovers whether sensitivity to expected ratio incorporates the likelihood of benefit.

13.1. Study 4 Materials and participants

Study 4 examined the effect of probability across the same expected ratios as Study 3: 1v5 (ER = 5), 1v2 (ER = 2), and 4v5 (ER = 1.25).

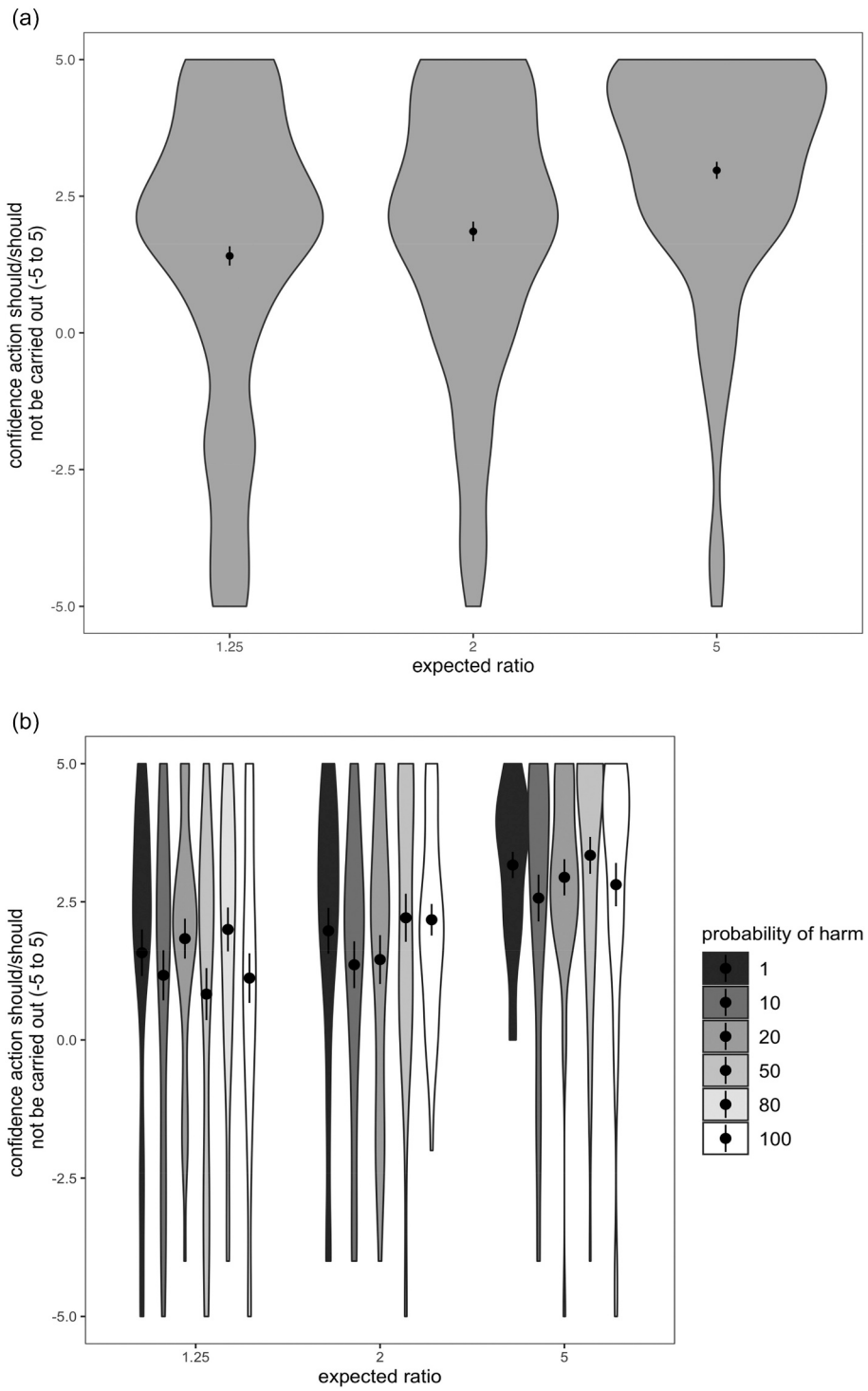

Fig. 3. a and b Sensitivity to expected ratio of lives saved to lives lost when rating confidence in action (3a). Insensitivity to variations in the probability of harm for each expected ratio (3b). Error bars represent one standard error. 80% harm only tested for 4v5 because other ratios cannot achieve it with whole numbers.

While the expected ratio of the number of lives being saved was held constant, probabilities of the missile hitting the group on the saving side were systematically varied: the missile had a 1%, 10%, 20%, 50%, or 100% chance of hitting the group of people the agent was considering saving, *see* Table 3 *for scenarios*. Participants were asked, for example, *Should Bob set off a rocket that he knows will kill 1 person, but that he also knows will destroy a missile that will otherwise have an X % chance of killing Y people* [where XY = 2]? (−5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). 697 participants were recruited (603 passed an attention check and were retained for analysis (86.5% passed); mean age = 34.3, SD = 11.5, 57.7% identified as female; 42.3% as male). Sample size was determined prior to data collection, and was selected to be able to detect a small-medium effect of probability, which was determined to require >550 participants (f = 0.12, alpha = 0.05, beta = 80%, two-tailed) using Gpower software (Faul et al., 2007). Sensitivity power analyses were conducted for Study 1 (using alpha = 0.05 and beta = 80%, two-tailed) to determine a minimum detectable effect size (MDES) of f = 0.11, using Gpower software (Faul et al., 2007).

Table 2
Study 3 Results.

EV Ratio	Probability	Mean	SD	% who would act
1.25	1	1.58	2.82	78%
1.25	10	1.17	2.91	71%
1.25	20	1.83	2.17	81%
1.25	50	0.83	3.02	68%
1.25	80	2.00	2.45	74%
1.25	100	1.12	2.91	64%
2	1	1.98	2.63	80%
2	10	1.36	2.54	69%
2	20	1.45	2.54	67%
2	50	2.21	2.68	74%
2	100	2.18	1.81	80%
5	1	3.17	1.42	92%
5	10	2.57	2.58	81%
5	20	2.94	1.94	91%
5	50	3.34	2.14	90%
5	100	2.81	2.39	86%

Table 3
Study 4 scenarios.

ER	Probability of saving	Scenarios
1.25	1%	Kill 4 people to 1% chance of saving 500 others
1.25	10%	Kill 4 people to 10% chance of saving 50 others
1.25	20%	Kill 4 people to 20% chance of saving 25 others
1.25	50%	Kill 4 people to 50% chance of saving 10 others
1.25	100%	Kill 4 people to 100% chance of saving 5 others
2	1%	Kill 1 person to 1% chance of saving 200 others
2	10%	Kill 1 person to 10% chance of saving 20 others
2	20%	Kill 1 person to 20% chance of saving 10 others
2	50%	Kill 1 person to 50% chance of saving 4 others
2	100%	Kill 1 person to 100% chance of saving 2 others
5	1%	Kill 1 person to 1% chance of saving 500 others
5	10%	Kill 1 person to 10% chance of saving 50 others
5	20%	Kill 1 person to 20% chance of saving 25 others
5	50%	Kill 1 person to 50% chance of saving 10 others
5	100%	Kill 1 person to 100% chance of saving 5 others

13.2. Study 4 Results

As in Study 3, a two-way ANOVA with probability and expected ratio entered as categorical variables revealed a significant effect of expected ratio, $F(2, 588) = 3.87$, $p = .021$, $\eta_p^2 = 0.013$ (mean 4v5 = 0.632, SD = 3.03; mean 1v2 = 1.12, SD = 3.11; mean 1v5 = 1.42, SD = 3.00); see Fig. 4a. A planned contrast revealed a linear effect of ER, $t(588) = 2.67$, $p = .008$, $r = 0.11$, but planned pairwise contrasts between expected ratio conditions show mixed results: the difference between ER = 2 and ER = 5 was not significant, (Welch t(396.01) = −1.003, Holm adjusted $p = .317$, $d = 0.10$), nor was the difference between ER = 2 and ER = 1.2 (Welch t(398.76) = 1.580, Holm adjusted $p = .115$, $d = 0.15$), but the difference between ER = 5 and ER = 1.25 was (Welch t(402.99) = 2.638, Holm adjusted $p = .026$, $d = 0.26$). Unlike situations involving probabilities of harming, there was a significant effect of probability of saving while keeping expected ratio matched, $F(4, 588) = 17.3$, $p < .001$, $\eta_p^2 = 0.105$, (mean 1% = −0.41, SD = 3.47; mean 10% = 0.739, SD = 3.20; mean 20% = 1.02, SD = 2.65; mean 50% = 1.23, SD = 2.92; mean 100% = 2.62, SD =2.15). A planned contrast revealed a linear effect of probability, $t(588) = 7.89$, $p < .001$, $r = 0.31$. The pattern is consistent with participants being risk averse to probabilistic saving: they were more likely to endorse the action when the chances of saving the group was high (e.g., 100%) and less likely to endorse it when the chances were low (e.g., 1%), even though the expected ratio was constant; see Fig. 4b. The interaction of probability and expected ratio was not significant $F(8, 588) = 1.49$, $p = .156$, $\eta_p^2 = 0.020$ (see Table 4). A logistic regression with expected ratio and probability entered as categorical variables confirmed that binary "should act" / "should not act" decisions were also sensitive to expected ratio, $\chi^2(2, N = 603) = 8.62$, $p < .001$ (proportions in favor of acting: 0.706, 95% CI = [0.640,0.765] when ER = 5; 0.682, 95% CI = [0.614,0.743] when ER = 2; 0.525 [0.456,0.592] when ER = 1.24); A planned contrast revealed a linear effect of ER, OR = 1.93, 95% CI [1.40, 2.70]. Planned pairwise contrasts between expected ratio conditions revealed a significant difference between ER = 5 and ER = 1.25, OR = 2.182 95% CI = [1.449,3.286]; as well as between ER = 2 and ER = 1.25, OR = 1.943, 95% CI = [1.294,2.916]; but not between ER = 5 and ER = 2, OR = 1.123 95% CI = [0.733,1.720]. Binary decisions were also sensitive to probability of saving, $\chi^2(5, N = 603) = 56.5$, $p < .001$ (proportions in favor of acting: 0.400 95% CI [0.317,0.490] when probability = 1; 0.600 95% CI [0.508,0.685] when probability = 10; 0.658 95% CI [0.569,0.737] when probability = 20, 0.675 95% CI [0.587,0.751] when probability = 50, and 0.840 95% CI [0.765,0.894] when probability = 100). Planned contrasts revealed a linear effect of probability, OR = 1.546 95% CI = [1.362, 1.754]. The interaction of probability and expected value was not significant, $\chi^2(8, N = 603) = 10.3$, $p = .241$, see Table 4.

13.3. Study 4 Discussion

Participants continued to exhibit sensitivity to expected ratio, being more confident of the action's rightness as the expected ratio of the number saved to the number killed increased. However, in contrast to the reactions to uncertain harming, we found a significant sensitivity to probability on the saving side. Participants were averse to versions of plans that, though holding expected value fixed, probabilistically save lives. For example, when it came to sacrificing four to save an expected value of five, people were generally favorable when the saving of five was certain, and unfavorable when it was presented as a 1% chance of saving 500.

14. General discussion

Collectively, our studies show that people are sensitive to expected ratio in moral dilemmas, and that they show this sensitivity across a range of probabilities. The particular kind of sensitivity to expected value participants display is consistent with the view that people's moral judgments are based on one single principle of threshold deontology. If one examines only participants' reactions to a single dilemma with a given ratio, one might naturally tend to sort participants' judgments into consequentialists (the ones who condone killing to save others) or non-consequentialists (the ones who do not). But this can be misleading, as is shown by the result that the number of participants who make judgments consistent with consequentialism in a scenario with ratio of 5:1 decreases when the ratio decreases (as if a larger number of people endorse deontological principles under this lower ratio). The fact that participants make some judgments that are consistent with consequentialism does not entail that these judgments are expressive of a generally consequentialist moral theory. When the larger set of judgments is taken into account, the only theory with which they are consistent is threshold deontology. On this theory, there is a general deontological constraint against killing, but this constraint is overridden when the consequences of inaction are bad enough. The variability across participants suggests that participants have different thresholds of the ratio at which the consequences count as "bad enough" for switching from supporting inaction to supporting action. This is consistent with the wide literature showing that participants' judgments can shift within the same ratio, depending on, for example, how the death of the one is caused.

Making the harms of action uncertain has a limited effect on participants' moral choices. Participants' confidence about the moral rightness or wrongness of killing one to save five was no different from their confidence about the moral rightness or wrongness of subjecting one hundred people to a 1% risk of death to save five. This highlights the possibility that people judge in accordance with the same moral principles when it comes to harming others versus putting them at even slight risk of harm. It also indicates that, in this sort of dilemma at least,

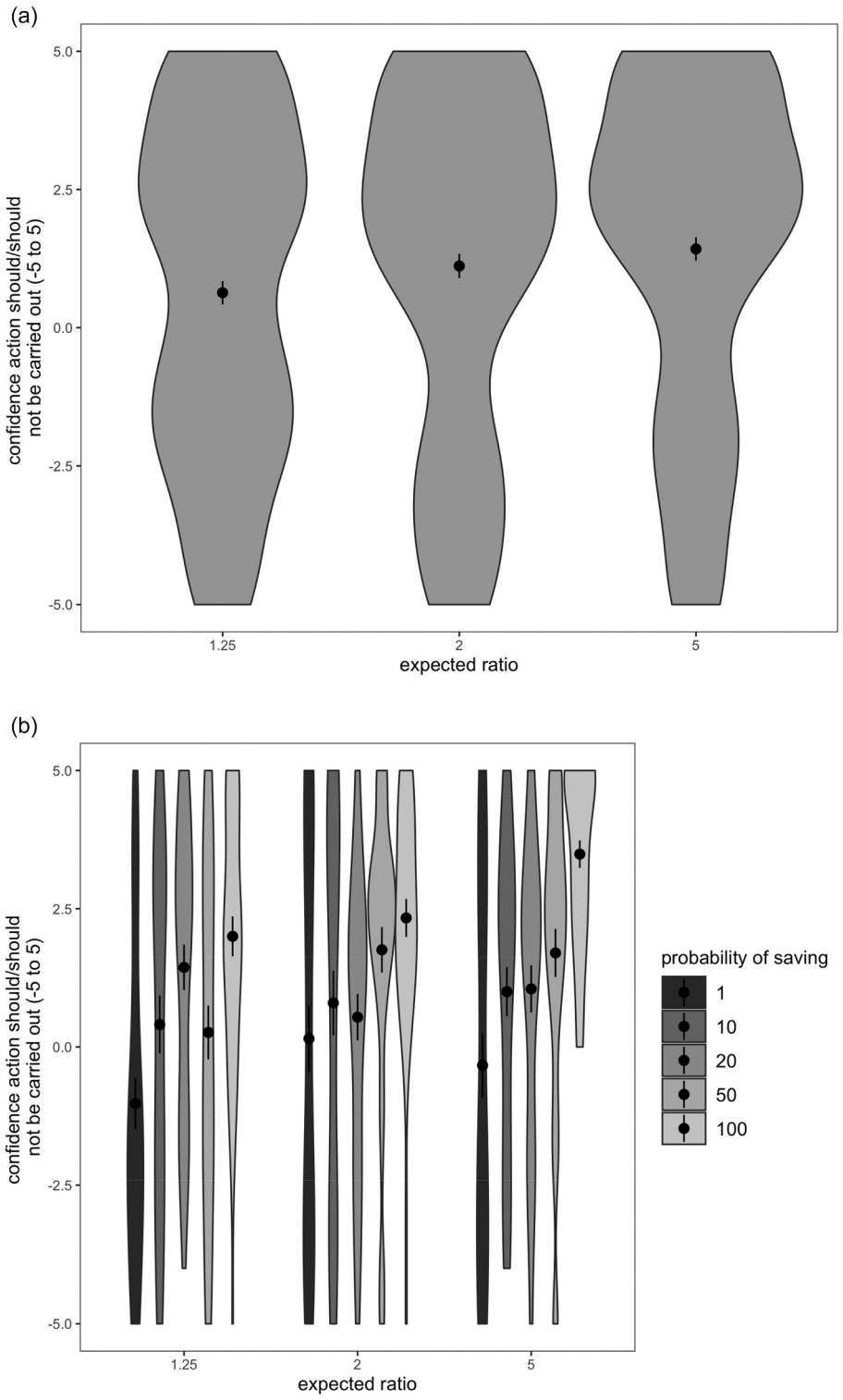

Fig. 4. a and b Sensitivity to expected ratio of lives saved to lives lost in more concrete scenarios in ratings of confidence in action when probability of saving is varied (collapsed across various forms of the same expected ratio; 4a). Sensitivity to various forms of the same expected ratio (4b). Error bars represent one standard error.

people do not show the usual risk-seeking tendency when it comes to losses.

Sensitivity to probability on the saving side revealed a somewhat different effect. In these cases, participants are less confident about the rightness of bringing about a particular level of harm when the benefit is uncertain, even when it has the same expected value. This is consistent with prospect theory and risk aversion for gains. This preference for concentrating benefits is sufficiently strong that participants are willing to take about half as much expected good for a given harm if that good is certain rather than probabilistic.

One possible explanation for the asymmetry between responses on the saving side and responses on the harm side might be that participants are bringing consistent deontological principles to bear. According to some deontological principles, such as DDA, one's duty not to kill is

Table 4
Study 4 Results.

EV Ratio	Probability	Mean	SD	% who would act
4v5	1	−1.02	2.96	27%
4v5	10	0.41	3.18	51%
4v5	20	1.44	2.63	66%
4v5	50	0.26	3.15	45%
4v5	100	2.00	2.37	72%
1v2	1	0.15	3.78	50%
1v2	10	0.79	3.64	62%
1v2	20	0.54	2.62	62%
1v2	50	1.76	2.63	80%
1v2	100	2.33	2.14	87%
1v5	1	−0.33	3.62	44%
1v5	10	1.00	2.77	67%
1v5	20	1.05	2.70	70%
1v5	50	1.70	2.76	78%
1v5	100	3.49	1.64	93%

stronger than one's duty to save or otherwise benefit others. Further, on such theories, one's duty to save or benefit others might be such that one can choose among a wide range of ways to fulfill the duty and, in some situations, there may be no duty to save or benefit others at all. (See the distinction between perfect and imperfect duties in Kant, 1785/2002) The scenarios are complicated in that they involve both imposing and reducing risk, but it is possible that in the case of benefiting others, given that there is no duty to benefit (or to benefit in any particular way) in the first place, one has no duty to distribute increased chances of living to more people as opposed to increasing even more the chances of living for a smaller group. Thus, with no such duty involved, but with a high value on certain saving, it makes sense in this case to prefer to save a smaller number with certainty than to perform an act that at best will decrease others' chances of dying when they might not have died in any case. And this is what we find in Study 4. In contrast, when we vary whether the agent would cause certain death or merely risk death, as we do in Study 3, we do not find a difference in participants' responses.

Our findings also contribute to a more nuanced understanding of deontology by comparing how participants respond to probabilistic and certain death. While deontologists may not be willing to kill one to save five, they may deem it acceptable to risk a 1% chance of harm to one to save five. Our data make salient the possibility that expected value calculation, rather than level of risk itself, accounts for this shift in judgment. An open question remains as to what determines a person's deontological weighting, or the value by which their expected value calculation is offset, in deciding whether to act.

Cohen and Ahn (2016) have proposed an alternative theory to explain people's moral judgments, namely, subjective utilitarianism, as a single process underlying people's judgments in moral dilemmas. The theory states that people choose the option that brings the maximal amount of personal value, with personal value purposefully left underspecified (Cohen & Ahn, 2016). It is possible that even if threshold deontology is the correct moral theory, it is not what is operative in actual moral decision-making. While a full comparative evaluation between subjective utilitarianism and threshold deontology is not possible here, we believe, on the basis of the studies above, that threshold deontology can better explain a persisting and quite systematic asymmetry in participant responses regarding doing and merely allowing harm.

Similar patterns of sensitivity to expected ratio and probability emerge in our findings with both concrete and abstract scenarios. The patterns also suggest that participants are inclined to make more extreme moral judgments (e.g., being more confident that it is morally acceptable to kill one in order to save two) in our concrete scenarios than they are in our abstract scenarios. This difference is consistent with some recent work on Construal Level Theory (CLT) and moral judgment (Gong & Medin, 2012; Lammers, 2012). Based on CLT, one possible explanation of the difference in our findings is that people engage in low-level construals in concrete scenarios, and such low-level construals can intensify moral judgments by being easier to imagine (see Gong & Medin, 2012, p. 635). By contrast, people engage in high-level construals in abstract scenarios, and these high-level construals involve greater psychological distance that can mitigate the extremeness of moral judgments. But it is not obvious how to apply the theory in this case, since participants are being asked to imagine both the possibility of two people dying and one person being killed intentionally by another. Since both aspects are made more vivid in the concrete scenario, it is not clear in which direction the moral judgment in this case we should expect to be intensified. We believe that the question of how responses differ with respect to concrete and abstract scenarios is an interesting one worth further exploration.

Our data suggest that people seem, on the whole, not to embrace simple consequentialist or absolutist non-consequentialist moral positions, but hold instead more nuanced views, balancing the harm done, the good achieved, and the value of rights, consistent with a principled threshold deontology. We are mindful of the fact that our data do not by themselves establish that such views are causally responsible for participant judgments in sacrificial moral dilemmas: further research will be required to establish whether threshold deontology is the main feature driving the psychological process of moral decision-making. But the data are consistent with the hypothesis that people's judgments are grounded in threshold deontology, rather than in consequentialism or absolutist deontology.

Our data also begin to shed light on the largely neglected domain of moral principles applied in an uncertain world. The normative ethical positions are largely silent on how such applications should be made, and so, given that almost every actual dilemma is likely to feature some degree of uncertainty at some level, data on how participants view such dilemmas is especially valuable and potentially relevant to social policies and procedures. In order to give herself a high probability of saving a small group of people (or a low probability of saving a large group of people), a firefighter might need to break a window that will cause a fire to reach an elderly person who is unable to move. What should she do? Should an autonomous vehicle be programmed to avoid plowing into a school bus by moving to the left, where there is a low probability of colliding with a tandem, or by moving to the right, where there is a high probability of colliding with a pedestrian? Among the factors that appear to be important, and worthy of serious further scrutiny, are whether the uncertainty is on the harm side or the benefit side, and whether the dilemma is about whether to incur that harm or instead how to apportion it.

Declaration of Competing Interest

The authors declare no conflict of interest.

Data availability

The data that support the findings of this study are openly available at https://osf.io/ftsem/?view_only=483d5e141b414ad2b392ff56b527e0ab

Acknowledgements

This research was supported by a National Science Foundation grant (SES-2049935) to C.R.M.M. The authors thank Anthony Gamst for his statistical advice.

References

Agerström, J., & Björklund, F. (2009). Moral concerns are greater for temporally distant events and are moderated by value strength. *Social Cognition, 27*(2), 261–282.
Alexander, L., & Moore, M. (2016). Deontological ethics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/. Retrieved from.

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23*(8), 861–868.

Arvan, M. (2013). Bad news for conservatives? Moral judgments and the Dark Triad personality traits: A correlational study. *Neuroethics, 6*(2), 307–318.

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 108*(2), 381–417.

Brennan, S. (1995). Thresholds for rights. *Southern Journal of Philosophy, 33*, 143–168.

Cohen, D. J., & Ahn, M. (2016). A subjective utilitarian theory of moral judgment. *Journal of Experimental Psychology: General, 145*(10), 1359.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of Personality and Social Psychology, 104*(2), 216.

Conway, P., Goldstein-Greenwood, J., Polacek, D., & Greene, J. D. (2018). Sacrificial utilitarian judgments do reflect concern for the greater good: Clarification via process dissociation and the judgments of philosophers. *Cognition, 179*, 241–265.

Costa, A., Foucart, A., Hayakawa, S., Aparici, M., Apesteguia, J., Heafner, J., & Keysar, B. (2014). Your morals depend on language. *PLoS One, 9*(4), Article e94842.

Costa-Lopes, R., Mata, A., & Mendonça, C. (2021). Real people or mere numbers? The influence of kill-save ratios and identifiability on moral judgements (¿Personas reales o meros números? La influencia de la proporción vidas sacrificadas/vidas salvadas y la identificabilidad en los juicios morales). *International Journal of Social Psychology, 36*(2), 378–395.

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

Cushman, F. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F., & Greene, J. D. (2011). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience, 7*(3), 269–279.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075.

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

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

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2018). Moral goodness is the essence of personal identity. *Trends in Cognitive Sciences, 22*(9), 739–740.

De Freitas, J., DeScioli, P., Nemirow, J., Massenkoff, M., & Pinker, S. (2017). Kill or die: Moral judgment alters linguistic coding of causality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(8), 1173.

De Freitas, J., & Johnson, S. G. (2018). Optimality bias in moral judgment. *Journal of Experimental Social Psychology, 79*, 149–163.

Diederich, A., Wyszynski, M., & Ritov, I. (2018). Moderators of framing effects in variations of the Asian disease problem: Time constraint, need, and disease type. *Judgment and Decision making, 13*(6), 529–546.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191.

Fitzpatrick, W. (2009). Thomson's turnabout on the trolley. *Analysis, 69*, 636–643.

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

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

Foot, P. (1984). Killing and letting die. In *Killing and letting die* (pp. 280–289). Fordham University Press.

Fried, C. (1978). *Right and wrong*. Harvard University Press.

Gong, H., & Medin, D. L. (2012). Construal levels and moral judgment: Some complications. *Judgment and Decision making, 7*(5), 628–638.

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

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

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.

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

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

Hansson, S. O. (2003). Ethical criteria of risk acceptance. *Erkenntnis, 59*(3), 291–309.

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

Hutcherson, C. A., Montaser-Kouhsari, L., Woodward, J., & Rangel, A. (2015). Emotional and utilitarian appraisals of moral dilemmas are encoded in separate areas and integrated in ventromedial prefrontal cortex. *Journal of Neuroscience, 35*(36), 12593–12605.

Kagan, S. (1989). *The limits of morality*. Oxford University Press.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Kamm, F. M. (1996). Morality, mortality. In , *vol. II. Rights, duties, and status*. Oxford University Press.

Kant, I. (1785/2002). *Groundwork for the metaphysics of morals*. Oxford University Press.

Körner, A., & Volk, S. (2014). Concrete and abstract ways to deontology: Cognitive capacity moderates construal level effects on moral judgments. *Journal of Experimental Social Psychology, 55*, 139–145.

Lammers, J. (2012). Abstraction increases hypocrisy. *Journal of Experimental Social Psychology, 48*(2), 475–480.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

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

Moore, M. (1997). *Placing blame: A theory of the criminal law*. Oxford University Press.

Nagel, T. (1979). War and massacre. In *Mortal Questions* (pp. 53–74). Cambridge University Press.

Nelkin, D. K., & Rickless, S. C. (2014). Three cheers for double effect. *Philosophy and Phenomenological Research, 89*(1), 125–158.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*(3), 530–542.

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

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

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.

Quinn, W. (1989a). Actions, intentions, and consequences: The doctrine of doing and allowing. *Philosophical Review, 98*(3), 287–312.

Quinn, W. (1989b). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy and Public Affairs, 18*(4), 334–351.

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

Rickless, S. C. (1997). The doctrine of doing and allowing. *Philosophical Review, 106*(4), 555–575.

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*, 38–68.

Ryazanov, A. A., Wang, S. T., Rickless, S. C., McKenzie, C. R. M., & Nelkin, D. K. (2021). Sensitivity to shifts in probability of harm and benefit in moral dilemmas. *Cognition, 209*, Article 104548.

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

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.

Sinnott-Armstrong, W. (2015). Consequentialism. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*. Retrieved from https://plato.stanford.edu/archives/win2015/entries/consequentialism/. Retrieved from.

Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology, 4*, 1–8.

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

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*, 1395–1415.

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

Thomson, J. J. (2008). Turning the trolley. *Philosophy and Public Affairs, 36*(4), 359–374.

Trémolière, B., & Bonnefon, J.-F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40*(7), 923–930.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458.