To appear in F. Aguiar, H. Viciana, & A. Gaitan (Eds.), *Issues in experimental moral philosophy*

Trolley Problems Reimagined: Sensitivity to Ratio, Risk, and Comparisons

Dana Kay Nelkin[1], Craig R. M. McKenzie[2,3], Samuel C. Rickless[1] and Arseny A. Ryazanov[3]

[1]Department of Philosophy, UC San Diego, [2]Rady School of Management, UC San Diego, [3]Department of Psychology, UC San Diego

Author note: This research was supported by a National Science Foundation grant (SES-2049935) to C.R.M.M.

Abstract: In this paper, we consider two ways in which traditional approaches to testing lay moral theories have oversimplified our picture of moral psychology. Based on thought experiments (e.g., Foot 1967 and Thomson 1976) concerning the moral permissibility of certainly killing one to certainly saving five, psychological experiments (e.g., Cushman et al. 2006) have been constructed that purport to sort instances of reasoning into an implicit consequentialist moral theory (according to which only consequences are morally relevant) or a non-consequentialist moral theory (according to which other considerations, such as rights, always override the maximization of good consequences). In earlier work (Ryazanov et al. 2020, unpublished manuscript, and in preparation), we have shown that asking participants questions in which we vary the ratio of lives saved to lives lost and probability of outcomes reveals that people may be appealing to a more subtle non-consequentialist theory known as threshold deontology. According to that theory, rights matter and can override the maximization of consequences in moral decision-making, but they do not *always* do so. Rights are not absolute in that if the cost to well-being becomes high *enough*, rights do not play an overriding role. We explain why varying the questions in these ways leads to a more nuanced and truer picture, and we briefly explore the implications of these results here. We then turn to a second way in which asking a different kind of question can bring out a fuller picture of implicit moral theorizing. We discuss several studies that ask participants *comparative* questions that involve not just comparing acting in such a way that there is a risk of killing vs. allowing to die, but how the risk is distributed. We find that asking the comparative question offers intriguing results that contrast with those that emerge from asking non-comparative questions. We then consider the challenging methodological question of whether we should privilege results based on one way of asking the questions over the other.

1. Introduction

Consequentialism is an elegant moral theory. 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).[1] In deciding which action to adopt, the consequentialist looks to the total value of the outcome of each action, a result

[1] For simplicity of presentation, here and below, we use "permissible" and "required" to mean "morally permissible" and "morally required" respectively.

of weighing the harms and benefits.

Non-consequentialists deny that consequentialism is true. This does not mean that they take consequences to be irrelevant to what we should and should not do; it simply means that they recognize other relevant considerations. For example, many non-consequentialists take it that *how* consequences are brought about can matter to what people ought to do. 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). And in turn, this is explained by the fact that we have rights against everyone that they not harm us in certain ways, but not rights against everyone that they assist us. Or consider 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., Quinn 1989a; Quinn 1989b; Rickless 1997; Nelkin & Rickless 2014). In turn, this might be explained by the fact that we have special rights not to be used without our consent.

The most influential way of trying to adjudicate between the consequentialist and non-consequentialist views has been by appeal to thought experiments, assessed through the method of reflective equilibrium (Rawls 1970; Daniels 2016). The method works by eliciting intuitions or responses to particular scenarios and then balancing these with general moral principles that can *explain* many of our intuitions to particular scenarios and that are themselves plausible. We might not be able to preserve all of our initial intuitions to particular cases where they are not consistent with each other and the general plausible principles with most explanatory power. So, intuitions about particular cases do not by themselves determine the correct moral theory. But they are important input to the balancing that allows us to reach reflective equilibrium.

Beginning with the groundbreaking work of Foot (1967) and Thomson (1976), the field has been dominated by a series of thought experiments in which we hold fixed a balance of harms and benefits and then vary considerations distinct from consequences. Start with a rescue case in which one is forced to choose between rescuing one person and rescuing five (otherwise similar) people. It is intuitive 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), thereby providing motivation for consequentialism. But non-consequentialists do not deny that consequences are morally relevant; they question whether they are the *only* morally relevant consideration. So, nonconsequentialists have appealed to other cases with the same balance of gain (typically, five lives) and loss (typically one life), but which introduce additional factors. 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, 2008; Fitzpatrick 2009), and fatally harvesting all of the organs of a healthy patient in order to transplant them into five patients who will die without the appropriate organ transplant (Foot 1967; Thomson 1976). It has seemed to many that acting in a way that causes the one to die in these cases is

impermissible despite the consequence that five would be saved, and such cases have been thought to support non-consequentialism.

But this has not ended the debate between consequentialists and non-consequentialists. One reason for this is that such intuitions can be tested systematically, and indeed psychologists and philosophers engaged in this project have produced a number of interesting findings (Cushman, Young, & Hauser, 2006; Hauser, Cushman, Young, Kang-Xing, & Mikhail, 2007; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Moore, Clark, & Kane, 2008; Arvan, 2013, De Freitas, DeScioli, Nemirow, Massenkoff, & Pinker, 2017). Our aim in this chapter (and in our experimental work, some of which we describe below) is to add to this rich body of work by showing how certain methodological assumptions and choices have constrained the scope of insights we might gain in engaging in this project.

In surveying the relevant philosophical and psychological literature we noted that much of the work presupposes a dichotomy between consequentialist reasoning on the one hand and a particular version of non-consequentialism, which we call "absolute deontology," on the other hand. According to absolute deontology, no amount of expected value could override moral principles that appeal to other considerations such as a prohibition on intending or doing harm, or violating others' rights.

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 nonconsequentialists have suggested that it is permissible to do or intend harm, or infringe others' rights, 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 absolute deontologist would say that it is 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).

In this chapter, we show how varying some common aspects of questions that participants are often asked to elicit evidence for their implicit moral theories produces responses that are consistent with a principled implicit threshold deontology. In doing so, we explain how some of our recent experimental findings undermine some common methodological assumptions, while at the same time they raise new questions for moral theorists. In section 2, we explain how the original and oft-repeated choice of scenarios that include a 5:1 ratio of those saved to those killed has prevented us from seeing certain *patterns* of responses that are consistent with a principled threshold deontological theory. In section 3, we show how changing the probabilities of both killing and saving in moral dilemmas elicits patterns of responses that are not consistent with the exclusive appeal to expected value predicted by a consequentialist theory. These results, too, as we will show, are consistent with a threshold deontological theory, although they expose gaps in the most well-developed current versions of such a theory. In section 4, we show how surprising differences emerge between separate and simultaneous evaluations of scenarios and explore some explanations

for this result. In section 5, we bring together a set of insights that emerge from this work and that underscore the valuable synergy between experimental and moral theoretical work.

2. Ratio, Difference, and Threshold Deontology

Asking participants the single question of whether it is permissible to kill one to save five allows for answers to one specific ratio of lives taken and saved. The resulting yes or no answers have been taken to map on to consequentialism and non-consequentialism respectively. But when we ask a variety of questions, the mapping onto moral theories can become more complicated. In particular, by varying the numbers of people saved, we leave open the possibility that people do not answer the questions in the same way. Suppose, for example, that participants were to agree that killing one to save five is permissible, but also believe that killing one to save *two* is *im*permissible? This pair of commitments would not be consistent with consequentialism, nor would it be consistent with absolute deontology. For on the one hand, the denial that it is permissible to kill one to save two departs from the consequentialist account that takes the overall value of consequences to be determinative in such moral matters, while the acceptance that it is permissible to kill one to save five departs from the absolute deontological account that prohibits any violation of rights no matter how valuable the consequences. Nevertheless, the pair of commitments is consistent with threshold deontology, as long as the relevant threshold is set somewhere between the 1:2 tradeoff and the 1:5 tradeoff. Now, very often threshold deontologists assume that the threshold is so high that only catastrophic losses can override the presumption that rights are to be respected (see, e.g., Nozick 1974 on "catastrophic moral horror," Walzer 1977 on "supreme emergency," and Moore 1997 on "extraordinary harms"—see also Nagel 1979 on "very large evils"). But the framework of threshold deontology does not by itself entail any particular threshold, and it is possible that people actually set a threshold at a much lower point. In fact, we believe that this is precisely what explains otherwise puzzling results.

In a series of experiments, we sought to examine just where participants might be setting thresholds, if indeed that is what they are doing. In fact, participants were very sensitive to the ratio of those killed to those saved.

In one experiment (Ryazanov et al., 2021),we presented participants with variations of the following scenario:

> *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, one would be killed to save one hundred, or 99 would be killed to save 100. 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).*[2]

The results are striking: for a case in which 99 could be killed to save 100, most participants were confident that the action should not be carried out; for a case in which one person could be killed to save two, participants were about evenly split; and for a case in which one could be killed to save 100, most participants were confident that the action should be carried out. (See figure 1.)

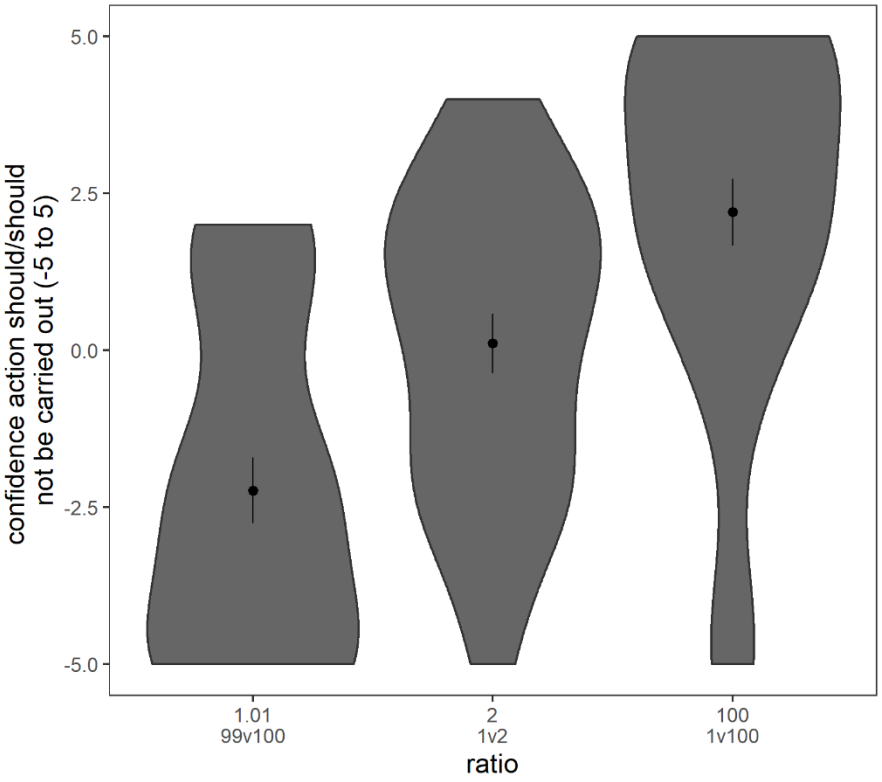

Figure 1. Sensitivity to ratio of lives saved to lives lost in ratings of confidence in action. Error bars represent one standard error.

Not only is there an increase in confidence as the ratio increases, but there is a change in average valence from negative to neutral to positive. This provides support for the idea that

[2] We chose this dependent variable, rather than a continuous forbidden-obligatory measure, because, according to many consequentialists and nonconsequentialists 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.

people are neither consequentialists nor absolute deontologists, but might instead reason consistently with threshold deontology.[3]

In theory, there are two ways that people might set the threshold for when positive consequences override respecting rights: by appeal to the *ratio* between number of people killed and those saved, as we have so far assumed, or by appeal to the *difference* between those two numbers. In fact, Figure 1 is not only consistent with people being sensitive to the ratio, it also shows that people are sensitive to it even when the difference remains the same. The first two conditions in Figure 1 hold the difference constant (99-100 vs. 1-2), but people are much less confident in carrying out the action in the first case where the ratio is lower. In a second study, we asked participants to respond to similar scenarios which kept ratio fixed and varied the difference: 1:2, 10:20, and 100:200. If the difference between the number of people killed and the number of people saved were operative, then we should expect significant variation in participant responses as the difference rises from 1 to 10 to 100. In fact, the mean answer for each was nearly identical (Ryazanov et al., unpublished manuscript).[4] While we do not know of extant threshold deontological theories that explicitly address whether the threshold should be set by ratio or difference, it is an interesting question whether such theories *should* incorporate a threshold-setting formula that appeals to ratio or one that appeals to difference or one that appeals to both.

In fact, there is reason to think participants are tracking the rational option here in privileging ratio over difference. The reason is that the trade-off between 100 lives ended and 200 lives saved seems to be equivalent to a series of pairwise trade-offs between one and two. While this suggestion clearly merits further exploration and the way people actually reason does not by itself entail a correct moral theory, we believe it is a good starting point for theorists to consider in filling a gap in their normative account.

---

[3] It is also worth considering subtle versions of consequentialism that purport to accommodate apparently deontological intuitions, such as a theory known as "rights consequentialism" (see Sinnott-Armstrong, 2019, section 5, for discussion). According to this theory, a death by killing (or by violating "rights") has more disvalue than a death by natural causes, other things equal, just as a bone fracture typically has more disvalue than a paper cut for purposes of a consequentialist cost-benefit calculation. While it might initially seem that such a consequentialist theory can also accommodate apparently "mixed" results for moral dilemma tasks as well as a threshold deontological theory, we believe that our particular studies are not in fact consistent with it. The reason is that in the cases we present, participants are asked to weigh the permissibility of killing one against the permissibility of allowing killings of two or five and more. This means that even according to the sophisticated consequentialist who recognizes great disvalue in death by killing, killing one will always have less disvalue relative to killing any number higher than one. Thus, rights consequentialism should treat killing one to save two from being killed and killing one to save five from being killed in the same way. Our findings are not consistent with the implicit operation of such a theory. We thank a reviewer for urging us to consider this kind of version of consequentialism.

[4] We also performed a similar study but instead of using an abstract scenario, we used a concrete one concerning the possible shooting down of a missile fired toward an inhabited house that would cause a smaller number to be killed. Consistent with other studies that show greater willingness to act when faced with concrete scenarios, the mean in response to all three scenarios (1:2, 10:20, 100:200) was higher than in the abstract versions, but what is most important here is that the mean response to each of the three concrete scenarios was again nearly identical.

By asking questions about moral permissibility that concern different ratios, we gain evidence that people are indeed operating with a threshold deontological picture. And by bringing this nuanced moral theory to bear on experimental design, we are able to construct opportunities to elicit just such evidence.

3. Ratio and Risk

Adding Probability to the Picture

So far, we have discussed scenarios in which it has been stipulated that harm is certain to come to some while others are certain to be saved. But in real life, we are rarely certain of anything.[5] How do people reason in these kinds of cases? Would 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?

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 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 for non-consequentialism in the form of either absolute or threshold deontology, where the value of outcomes does not entirely determine what one ought to do. For example, it is not obvious from the perspective of threshold deontology what are the contours of 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.

Our interest in examining ratio in the context of uncertainty of outcomes leads us to define expected ratio. We define 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:

[5] 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).

$EV_{lives\ saved} = N_{lives\ saved} \times P_{lives\ saved}$
$EV_{lives\ ended} = N_{lives\ ended} \times P_{lives\ ended}$

With these variables defined, we can define expected ratio:

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

With this in mind, we can more precisely formulate our question: is expected ratio the only factor that affects whether respect for rights should be overridden, or does probability play an independent role? To test this, we conducted a study in which we presented each participant with one of three scenarios, otherwise alike, but in which the expected ratios varied (ER=5, ER=2, and ER=1.25), and then varied each of these by changing the numbers and probabilities in each outcome while keeping the expected value fixed. For example, in the ER = 5 cases, 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*).

The findings were once again clear, with expected ratio alone accounting for the patterns of response: regardless of how the expected value of harm was presented to participants (e.g., 1% chance of killing 100 to save 5, or 1 certainly being killed to save 5), participants remained sensitive to the expected ratio. (See figure 2.)

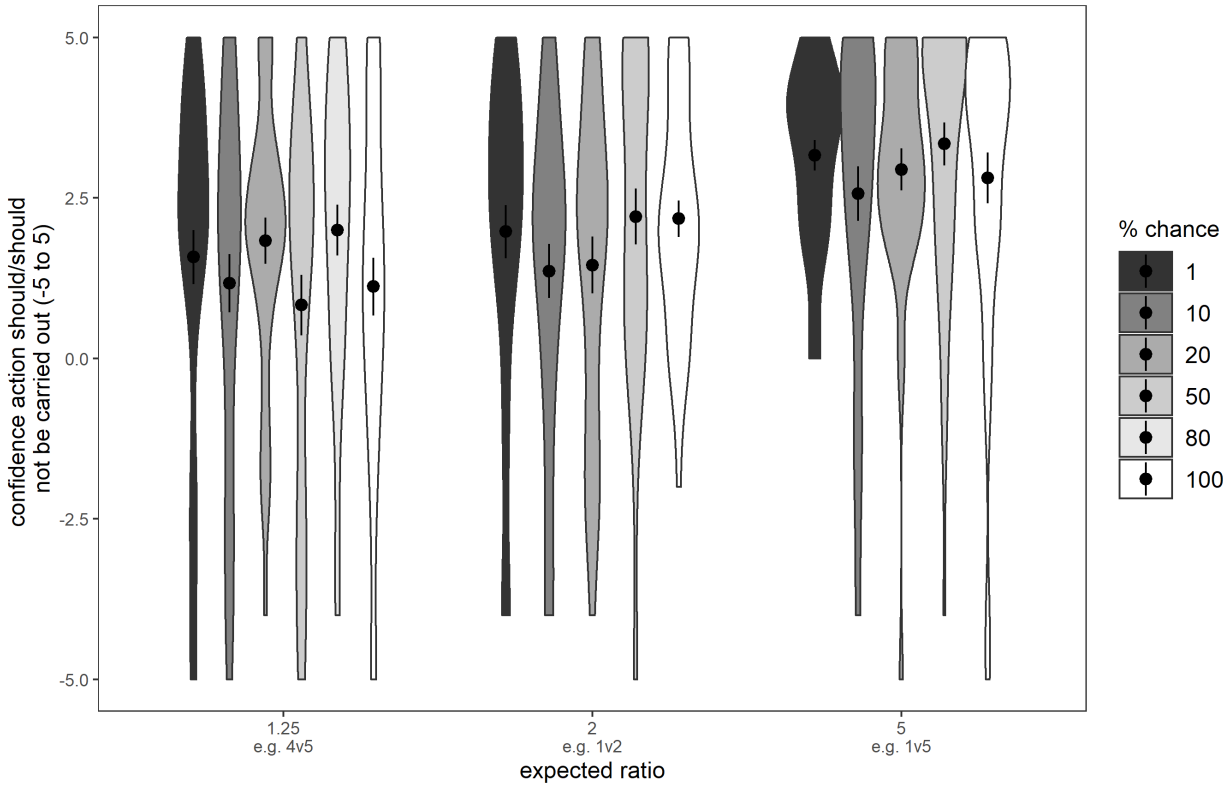

Figure 2. Sensitivity to expected ratio in ratings of confidence in action, but insensitivity (within each expected ratio) to the probability of harm, holding the expected value of harm fixed. Error bars represent one standard error. 80% harm tested only for 4v5 because other ratios cannot achieve it with whole numbers.

But this is not the end of the story. For when we instead kept fixed the certainty of harm along with expected ratio, and instead varied the probabilities of *saving*, we found very different responses. The question here is this: would an action that kills one to certainly save two be judged differently from an action that kills one to save four who have a 50% chance of dying without the intervention? In each case, the expected ratio is the same, but the probabilities of achieving the end of saving varies. In contrast to the earlier results, we found that participants were averse to versions of plans that, though holding expected value fixed, probabilistically save lives (Ryazanov et al unpublished manuscript). 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.[6] (See figure 3.)

[6] It is important to note that the confidence that one should act when sacrificing four to save five here is in some tension with the earlier result that showed only a neutral response to the question of whether one should act when sacrificing one to save two. At first glance, these seem inconsistent: the expected value in the first case is lower than that in the second, so it seems that, if anything, one should favor acting in the case where expected value is higher, not lower. We believe that this tension is due to the fact that there are systematically different

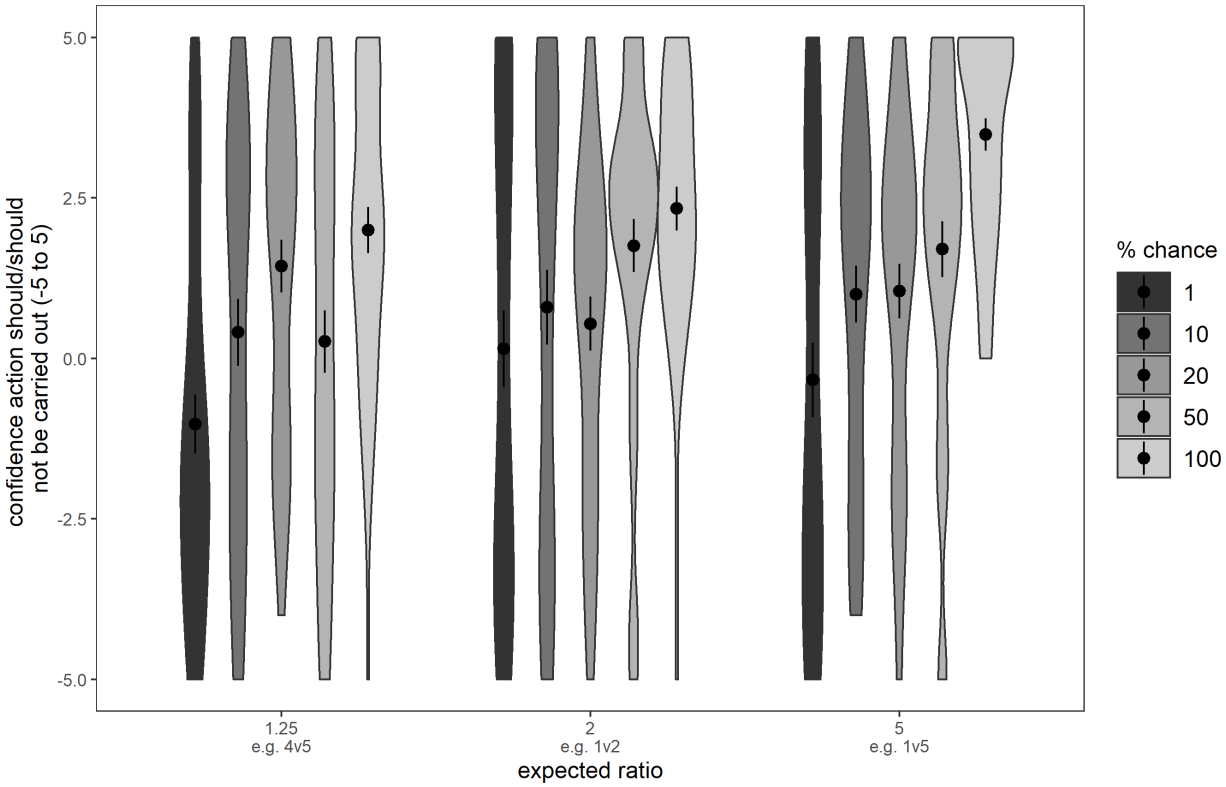

Figure 3. Sensitivity to probability but not expected ratio in ratings of confidence in action when probability of saving is varied. Error bars represent one standard error.

Is there a principled way to account for this asymmetry between the effects of probability on harming versus saving? Prospect theory, a theory of decision making under risk, posits that people are risk averse for gains (preferring sure gains to risky gains with the same expected value) and are risk seeking for losses (preferring risky losses to sure losses with the same expected value; Kahneman & Tversky, 1979). This is due to diminishing sensitivity to both increasing gains and increasing losses: Gaining twice as much money does not have twice the utility, and losing twice as much money does not have twice the disutility. Assuming that saving people is a gain and harming people is a loss, prospect theory can potentially explain the above results in that we find risk aversion for saving but not for harming; in the latter case, participants are risk neutral (neither risk averse nor risk seeking). Our finding of risk neutrality for harm is a bit puzzling from the perspective of prospect theory, given that risk seeking behavior for losses is usually robust, but nonetheless the decreased risk aversion for losses relative to gains is consistent with the theory. Interestingly, though, we note again that a deontological theory is consistent with the asymmetry. Indeed, it can potentially explain the

responses to abstract and concrete scenarios, a pattern that has been well-documented. (People are more inclined to favor acting when presented with concrete scenarios.) This does not remove the tension, of course, and it is something we hope to explore in more detail in the future. For now, it is worth noting that when we compare apples to apples (concrete to concrete and abstract to abstract) we are seeing clear patterns of response.

nature of the asymmetry for risky saving and risky harming, and do so in terms that are more natural and explanatory for moral reasoning. In particular, according to the Doctrine of Doing and Allowing, one's duty not to kill is stronger than one's duty to save or otherwise benefit others, and on some theories one has a variety of options as to how to fulfil one's duty to benefit others. Given that people place a high value on certain saving and that there is no particular duty to distribute chances of being saved across a larger number of people, it makes sense that some will take it that one ought to save a smaller number with certainty when another life will be ended in order to achieve the result. This hypothesis, of course, calls for more exploration by moral theorists and psychologists alike.[7]

Location of Shifts in Probabilities

In seeking a comprehensive account of how probability affects moral reasoning, it is also crucial to note that it is often simply assumed that the baseline probability of someone being killed or being saved is either 0 or 1, and that one's action can alter that starting point by raising it from 0 or lowering it from 1. But in real life things are once again not so simple. The starting probability of a person's being killed might be .5 and, by acting, one can expect to change that to any value between 0 or 1, such as .25 or .75 or .38. In Ryazanov et al. (2021), we set out to test whether the location on the scale between 0 and 1 of a shift in probability is a factor in moral reasoning.

If people are implicitly appealing to consequentialist principles, then they should be indifferent to fixed changes in the probability that they would cause death at different locations on the probability scale, since the expected value difference between acting and not acting will be identical. For example, there should be no difference in reaction to whether a person should increase someone else's risk of death from 0-25% or from 75-100%. Similarly, they should prefer smaller increasing shifts in probability of causing harm to larger ones in probability of causing harm. But in a series of experiments we found that participants did not respond in these ways. To the contrary, we found that participants are confident that, in order to save a group of 2 people, Harry should raise the probability of death for 4 bystanders from 0% to 25%, but that Harry should *not* raise the probability of death for a different 4 bystanders from 75% to 100%. See figure 4.

[7] The picture is also incomplete as it stands, in a way we bring out in section 4.

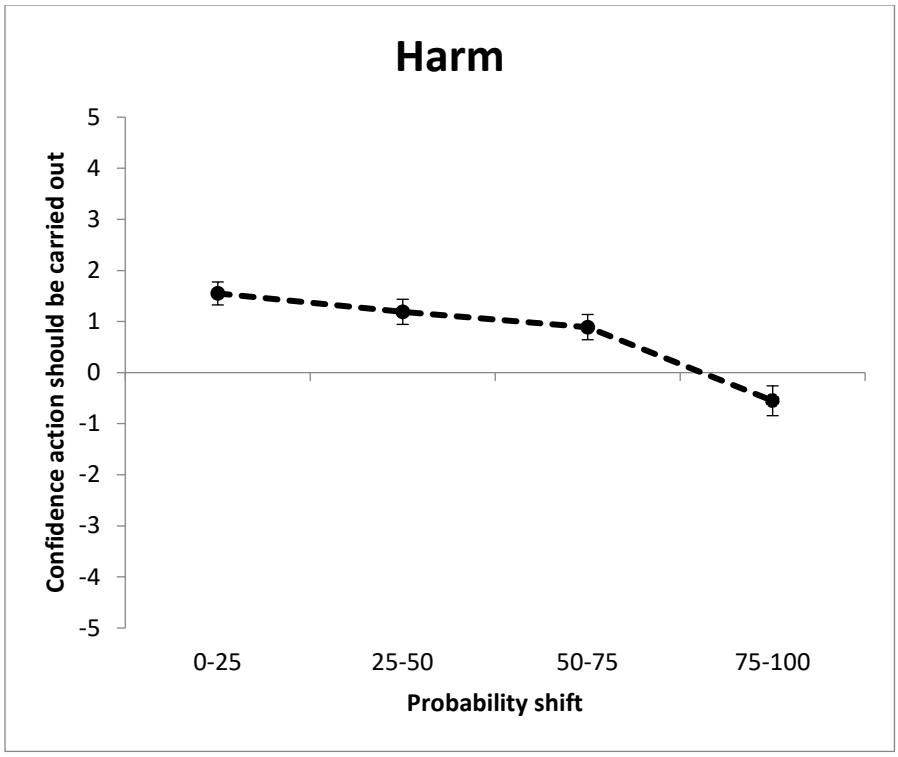

Figure 4. Sensitivity to probability shift location for an action that increases the probability of four bystanders dying by 25% in order to save two people. Error bars represent one standard error.

It seems to matter very much to participants what the end point is. This conclusion was underscored by the results of a further study in which we tested whether people would rather raise the probability of death to bystanders from 0 to x or from x to 100 for a number of values of x. The aim was to try to see more precisely how much of an expected value premium people think we should pay to avoid raising the end-point to 100%. When presented with the two options together of raising the probability of death for four bystanders from 0 to 95% or raising the probability of death for four bystanders from 95% to 100% in order to save two people, participants did not exhibit any significant preference for one option over the other. Participants were thus willing to pay a tremendous expected value premium to avoid raising the probability of death to 100%.[8]

What all this suggests is that people are not driven only by expected value, and that implicit consequentialist reasoning cannot account for these patterns of responses. At the same time, while the responses are consistent with implicit deontological reasoning, deontologists do not currently have any particular resources in their theories to explain them. Threshold deontologists would need to find more fine-grained moral principles than have been identified thus far. For example, it may be that the right against raising a person's risk of death

[8] A further study suggests that it is not only the endpoint of 100% that people wish to avoid. They also favor a shift from 5% to 50% over a shift from 50% to 95%, for example (Ryazanov et al., 2021).

to 100% is simply 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. But to our knowledge, these comparisons of stringency have not been investigated in any detail, and a theoretical rationale would need to be offered.[9]

4. Risk Considered Separately and Simultaneously

In a previous section, we noted that participants responded very similarly to each member of a pair of scenarios that kept expected value of harm fixed but varied the probability and numbers of those to whom harm was risked. Their mean rating of confidence in acting was nearly identical whether they were asked how confident they were that a person should impose a 1% risk of death on 100 to save five or how confident they were that a person should impose a 100% risk of death on one to save five. But, as has been shown in previous research, how people respond to the presentation of a single scenario in a between-subjects study (separate evaluation) does not perfectly predict how they will respond to the presentation of a choice between scenarios (simultaneous evaluation) (e.g., Hsee et al. 1996 and Hsee et al. 1999; see also Sher & McKenzie, 2014). When we sought to examine participants' responses to risk when choosing between the actions presented in two different scenarios, we also found that participants had a preference that was not captured by their responses to single scenarios. In particular, we found a significant preference for the idea that people should spread risk of harm across more individuals rather than concentrate a greater risk across fewer.

In Ryazanov et al. (in preparation), we asked participants to evaluate one of three scenarios, two of which were single scenarios of the sort described earlier, while a third asked participants to choose between the two plans presented in those scenarios. One plan was certain to kill one bystander, but would save two people from a missile that had accidentally been fired at them (e.g., 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 kill 2 people?; -5: *very confident he should not set off the explosion*, to 5: *very confident he should set off the explosion*). A second plan involved a 25% chance of killing four bystanders to save the two from the missile. The third scenario asked participants to choose between the two plans: *Assuming that Bob must fire one of the two rockets, which should he do: fire Rocket X, which he knows with certainty will both save the group of two people and at the same time kill one bystander; or fire Rocket Y, which he knows with certainty will both save the group of two people and has a 25% chance of killing 4 bystanders?* (Participants responded on a scale ranging from -5 = "*Very confident he should fire Rocket X" to 5 = "Very confident he should fire Rocket Y"*).

When participants were presented with a single plan and asked to express confidence in whether it should be carried out, their responses were again nearly identical. But when

[9] One starting point might be provided by a related study (Ryazanov et al., unpublished manuscript). There is some evidence that people perceive the shift with the higher endpoint as more harmful, and this perception partially mediates the bottom-line judgments of confidence in what agents should do. But whether this is a rational assessment, and, if so, why, is a further question.

presented with a choice between the two plans, there was a significant overall preference for the plan that spread the risk among four bystanders (imposing a chance of 25% that they would all die) to save two rather than certainly killing one to save two.[10] Thus, asking the comparative question brings out an aspect of implicit moral reasoning that is not captured by presenting only one scenario to each subject.

Interestingly, however, a similar experiment that varied the probabilities of saving rather than the probabilities of harming again exposed an asymmetry. As before, participants were more confident that one should harm one for certainly saving two than they were that one should harm one for a 25% chance of saving four. But, in contrast to the case in which the chance of harm varied, this preference for certain saving did not change when presented with the simultaneous evaluation and being asked to choose which of the two plans should be implemented.

Thus, two aspects of the results initially seem to call out for explanation: (i) the fact that a preference emerges only under simultaneous evaluation in the cases that vary according to the spread of risk of harm; (ii) the asymmetry between the patterns in the risk of harm cases and the risk of saving cases. Let us start with (i). It is notable that much past research on comparisons between separate evaluation and simultaneous evaluation concerns evaluative questions about either how good something is or how much one would pay for it. [Footnote with example?] In that case, there is a genuine puzzle when two options are evaluated as on a par separately but one is clearly preferred in simultaneous evaluation, because these evaluations appear inconsistent.[11] And there is an important question whether we should privilege the responses given separately or the ones given simultaneously as the participants' true or rational judgment.

But we believe that, in the case at hand, we need not choose one over the other. The reason is that the dimension of evaluation in this case is confidence in whether a plan *should* be carried out. And there is no contradiction or even tension between being very confident that Plan A should be carried out when the choice is between Plan A and not acting at all and being very confident that Plan B should be carried out instead of Plan A when that is the available alternative. At the same time, it is important to note that the simultaneous question that asks for a comparison of plans gives us new information that is simply not available when we ask the separate ones. The reverse is also true: we get information about what people think they

[10] The idea that spreading risk was preferred was further supported by experiments with the same general structure but which compared a greater total expected value of harm spread over more people to a lesser total expected value of harm spread over fewer. In these cases, participants became indifferent to which plan should be selected, which suggests that they are willing to pay a premium in expected value in order to spread risk across a larger number of people, thereby lowering each individual's risk of harm even though the total expected value of harm is higher (see Ryazanov et al. in preparation for details).

[11] There are also cases of complete "reversal" in which separately one object, A, receives a higher valuation than another object B, but when presented together there is a preference for B. There have been a number of interesting explanations of this phenomenon (see, for example, Hsee, 1996, Hsee et al., 1999, and Sher & McKenzie, 2014).

should do when the alternative is inaction. Thus, asking both kinds of questions gives us a more complete picture of people's implicit moral commitments.

The second aspect of the results poses a substantial challenge. Why should there be an asymmetry between the way people treat comparisons between plans that spread risk of harm differently and the way people treat comparisons between plans that spread the chance of saving differently?

While prospect theory predicts more risk seeking behavior for harm (losses) than for saving (gains), it is not clear why the risk seeking behavior only emerges when the harm scenarios are compared, and not when the scenarios are evaluated separately. It may again be that invoking deontological principles is better suited to explaining our pattern of results. According to this account, 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. On some deontological theories, such as ones that adopt the Doctrine of Doing and Allowing, one's duty not to kill is 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 participants' responses to separate evaluations. In contrast, when we vary whether the agent would cause certain death or merely risk death in separate evaluations, we do not find a difference in participants' responses. In that case, it is only in the simultaneous evaluation that we find differential responses. One reason for this might be that it is *pro tanto* wrong to impose significant risk, just as it is *pro tanto* wrong to kill, and so the salience of the comparison does not arise until one sees the two options in a simultaneous evaluation. Thus, while there may be no duty to save (or to save in a particular way or on particular occasions), and so it is permissible to take into account factors such as probability of having an effect, there is a duty at all times not to kill or impose significant risk. And this could go some way toward explaining why it is only once participants see a direct comparison of certain harming and spreading the risk of harm to a larger group that *this* difference becomes especially salient.

5. Putting It All Together And Looking Forward

We believe that the series of experiments we have described has positive implications for both future philosophical work and for future psychological investigation. First, considering threshold deontology as a serious contender in moral theory opens up the possibility of both

new experiments and new interpretations of data. We can see at least some sets of participant responses that might seem inconsistent against the backdrop of only the options of consequentialism and absolute deontology as perfectly rational on an implicit threshold deontological theory, for example. Importantly, this possibility only emerges fully when we reject the common assumption that the threshold is always to be set very high (e.g., when the fate of a city or the world is at stake), an assumption that the results of the experiments we describe give us some reason to question.

Further, considering the hypothesis that threshold deontology is implicit in our moral reasoning might have implications for some of the deeper explanations of our moral judgments in terms of particular psychological mechanisms. For example, and oversimplifying slightly, some have appealed to a dual systems account to explain seemingly inconsistent responses: we are using our fast, non-conscious system when we appeal to deontological principles, and we are using our slower, deliberative, conscious system when we calculate according to consequentialist principles (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene & Haidt, 2002; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene et al., 2009; Haidt, 2001; Cushman, et al., 2006; Cushman, Young, & Greene, 2010; Cushman & Greene, 2011; Paxton, Ungar, & Greene, 2012; Cushman, 2013; Crockett, 2013). If threshold deontology is operating as an implicit moral theory, then we have at least one reason to think that this division between systems is not playing a fundamental explanatory role. For it would be odd if we were to switch systems when faced with scenarios in which one must be killed to save two and in which one must be killed to save five. Rather, it seems that we are making nuanced assessments that take into account *at the same time and in a single calculation* both defeasible deontological principles and the consequences that can override them. We put this hypothesis forward tentatively, and at the same time offer it as an illustration of one general way in which philosophical theorizing can help generate questions for psychological research.[12]

[12] Others have provided different reasons for rejecting the dual process model as the best fit for explaining reactions to moral dilemmas. For example, Rosas et al. (2019) argue that the dual process model is committed to utilitarian reactions being associated with longer reaction times (since it is associated with the slower System 2), but that its predictions contradict the patterned reactions times when faced with cases on what they call a "utilitarian gradient". They construct a series of cases intending to increase the amount of utility gained by acting in each case while keeping fixed the violation of a deontological principle. For example, at one end, the benefits of acting are extremely high relative to the costs, and in the middle, the costs and benefits are relatively closer. Rosas et al find that at the extremes reaction times are faster than in the middle, and they posit that it is conflict of considerations in the middle of the gradient that causes the slowness in reaction times. We find these results intriguing and the conclusion congenial to our own. At the same time, we believe that their gradient of cases is constructed partly on the basis of considerations that are not in fact related to utility (e.g., whether the participant answering the question is described as one of the people to be saved) and partly on the basis of considerations that are deontological (e.g., whether the one to be killed is guilty of having threatened the five). One of their cases that clearly succeeds in isolating increased utility is a case in which one can save 100,000 instead of five by killing one. Their results with respect to this case are consistent with our own, while it is important to note that it is a "catastrophic" case to which we hope to have added by presenting cases in which much smaller numbers of lives are at stake (ee also Rosas et al.,(2021). Trémolière and Bonnefon (2014) also explore modifications to the dual process hypothesis to account for faster response times as cases include higher numbers of lives saved in moral dilemmas, even when the responses are "utilitarian". Their findings of increasing numbers of "utilitarian"

Moving in the other direction, the results of these experiments, like many others, promise to help fine-tune the intuitions in response to particular scenarios that provide input to reflective equilibrium. Equally importantly, they bring new questions to our attention. For example, how is the threshold in threshold deontology to be set? People seem to privilege ratio over difference in numbers of people harmed and saved. Is that as it should be? Given that intuitions are input to reflective equilibrium and not the final word on the normative questions of what is good or right or permissible, there remains work to do. But we have tried here to sketch a reason why our intuitions are tracking the normative truth. Similarly, we can ask whether probability provides an independent input to setting the threshold, and we have found different answers depending on whether the harm or the saving is probabilistic in comparison to certain. Moral principles, such as the Doctrine of Doing and Allowing and the Doctrine of Double Effect, are typically presented in non-probabilistic terms, but these results raise questions about whether we should recognize probabilistic versions of them and precisely what their content should be. For example, it seems plausible that the right against having the probability of one's death increased by another from 0% to 100% is more stringent than the right against having the probability of one's death increased by another from 0% to 5%. But participant responses do not suggest that they are operating with a linear function along the spectrum from 0% to 100% that corresponds to degrees of stringency. Are the responses tracking a normative truth in their patterns of response? These are deep and difficult questions, but given that in real life we are rarely certain of the outcomes of possible actions they are ones that call out for answers. We hope that the results and interpretations sketched here contribute to this project.[13]

References

Alexander, L., & Moore, M. (2016). Deontological ethics. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/win2016/entries/ethics-deontological/

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

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.

responses as the lives saved increases is consistent with our own findings, as well. In contrast to our approach, they are most focused on reaction times as a measure of effortfulness, and they do not attempt to reconcile the results with a possible consistent moral theory, instead conceiving of certain participant responses as "utilitarian" and others as "nonutilitarian".

[13] Many thanks to two reviewers for their very helpful comments and to the editors for the opportunity to contribute to this volume.

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.

Daniels, N. (2020). Reflective equilibrium. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/sum2020/entries/reflective-equilibrium/

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.

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

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

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.

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.

Kahenman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*, 263-292.

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.

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

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.

Rosas A., Bermúdez, J. P. & Aguilar-Pardo, D. (2019). Decision conflict drives reaction times and utilitarian responses in sacrificial dilemmas. *Judgment and Decision Making* 14 (5): 555 564.

Rosas, A., Bermúdez, J. P., Martínez Cotrina, J., Aguilar-Pardo, D., Caicedo Mera, J. C., & Aponte-Canencio, D. M. (2021). Perceiving utilitarian gradients: Heart rate variability and self-regulatory effort in the moral dilemma task. *Social neuroscience*, 16(4), 391–405.

Ryazanov, 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*.

Ryazanov, A., Wang, S. T., Nelkin, D. K., McKenzie, C. R. M., & Rickless, S. C. (unpublished manuscript). *Beyond killing one to save five: Sensitivity to probability and ratio in moral judgment*.

Ryazanov, A., Wang, S. T., Nelkin, D. K., Rickless, S. C. & McKenzie, C. R. M. (in preparation). *Moral dilemmas and the distribution of risk*.

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.

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.

Sher, S., & McKenzie, C. R. M. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General, 143*, 1127-1143.

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. (2019). Consequentialism. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), Retrieved from https://plato.stanford.edu/archives/win2015/entries/consequentialism/

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*, 204–17.

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

Walzer, M. (1977) *Just and unjust wars: A moral argument with historical illustrations*. New York: Basic Books.