Eliciting and Assessing our Moral Risk Preferences

Shang Long Yeo

(forthcoming in *American Philosophical Quarterly*)

Abstract: Suppose an agent is choosing between rescuing more people with a lower probability of success, and rescuing fewer with a higher probability of success. How should they choose? Our moral judgments about such cases are not well-studied, unlike the closely analogous non-moral preferences over monetary gambles. In this paper, I present an empirical study which aims to elicit the moral analogues of our risk preferences, and to assess whether one kind of evidence – concerning how they depend on outcome probabilities – can debunk them. I find significant heterogeneity in our moral risk preferences – in particular, moral risk-seeking and risk-neutrality are surprisingly popular. I also find that subjects' judgments aren't probability-dependent, thus providing an empirical defence against debunking arguments from probability dependence.

Keywords: ethics of risk; risk aversion; risk neutrality; risk seeking; debunking

We often find ourselves in situations of risk – our acts may only have some probability of harming or benefitting others, rather than doing so for sure. Any adequate moral theory must tell us how we should weigh the probabilities of benefits and harms in such scenarios. In the closely analogous domain of preferences over monetary gambles, we can identify three kinds of risk preferences: risk-aversion, on which we prefer receiving some money for sure over a bet that gives us the same expected return; risk-seeking, on which we prefer a bet over the certainty of receiving its expected return; and risk-neutrality, on which we are indifferent between a bet and the certainty of receiving its expected return. These risk preferences over monetary gambles, and the conditions under which they are rational, have been extensively studied in philosophy and beyond (Buchak 2013; Wakker 2010; Mata et al. 2018; Charness, Gneezy, and Imas 2013).

In contrast, much less attention has been paid to similar issues in the ethics of risk. How should we specify the moral analogues of the risk preferences? Which moral risk preference do we in fact adopt? Which risk preference *should* we adopt? In this paper, I hope to make headway on these questions with the help of an empirical study of our moral judgments about risky cases – where, for instance, we must choose between a rescue that saves more people with a low probability of success, and one that saves fewer people but is more likely to succeed. This study has two aims: first, to elicit our moral judgments about risky cases and to determine which moral risk preference is driving them, and second, to assess whether there is empirical evidence that undermines the reliability of such judgments. Here's how this paper will proceed: section 1 clears the ground for the elicitation of our moral risk preferences. Section 2 provides the background to evaluating the reliability of our moral judgments about risk. Sections 3 and 4 detail the results of my empirical study, and explores the implications for the ethics of risk. Section 5 concludes.

1. Our Moral Risk Preferences: What They Are and How to Elicit Them

In non-moral domains, risk preferences are clearly specified in terms of an agent's preferences over gambles – however, more work needs to be done to translate these into the moral domain. I start by defining the moral analogues of the risk preferences, with the help of a concrete case: suppose a captain is operating a ship in a furious storm, and they receive a call for help from a remote island where they find 50 people awaiting rescue. However, their ship is certified to carry less than 50 people, and they can attempt one of two options: a) a probabilistic rescue which carries a 90% probability of rescuing all 50, and a 10% probability of rescuing none (such that all 50 die), or b) a sure rescue which has a 100% probability of rescuing exactly 45 people (such that 5 die). (Assume that the sure rescue uses a lottery to pick which 45 will live and which 5 will die.) Morally speaking, how should they choose?

There are three possible verdicts here: either they're morally required to choose the sure rescue, or they're required to choose the probabilistic rescue, or they're permitted to choose either. These map on neatly to the non-moral risk preferences over money, which are defined in terms of preferences between a risky monetary gamble and the guarantee of receiving its expected return. For instance, an agent is strictly specifically risk-averse in money just in case for all x, they strictly prefer a guaranteed $x to a non-degenerate gamble whose expected return is $x (Buchak 2013, pp. 21–22). The moral analogue should similarly pick out the sure option from the set {sure option, probabilistic option}, so it's naturally understood as the moral *requirement* to choose a sure option that confers some benefit (or harm) of size x with certainty, over a probabilistic option that results in an expected benefit (harm) of size x. An agent who is risk-neutral in money is often understood as one who, for all x, is indifferent between a guaranteed $x and a risky gamble whose expected return is $x. The moral analogue would deem both the sure and risky options choiceworthy, so it maps onto the verdict that both options are morally *permissible*. Finally, if we also substitute strict preference for moral requirement with regards to moral risk-seeking, we get the following moral analogues:

> **Moral risk-aversion:** For all x, an agent is morally required to choose a guaranteed benefit (or harm) of size x over a probabilistic option whose expected benefit (or harm) is of size x.
>
> **Moral risk-seeking:** For all x, an agent is morally required to choose a probabilistic option whose expected benefit (or harm) is of size x over a guaranteed benefit (or harm) of size x.
>
> **Moral risk-neutrality:** For all x, an agent is morally permitted to choose a guaranteed benefit (or harm) of size x, or to choose a probabilistic option whose expected benefit (or harm) is of size x.[1]

[1] Buchak (2013) defends a rank-dependent expected utility theory on which different risk preferences are rationally permissible – for instance, we're rationally permitted to be risk-neutral, but also permitted to be risk-

In the rescue case, the probabilistic rescue confers an expected benefit of the same size[2] as the sure rescue's guaranteed benefit (45 people survive, 5 die). Moral risk-aversion thus implies that we are morally required to choose the sure rescue; moral risk-seeking implies we're required to choose the probabilistic rescue; moral risk-neutrality says we're permitted to choose either.

Different theoretical rationales have been offered in support of each risk preference. First, moral risk aversion might be supported by arguing that we don't know the people in this case and what their risk preferences are like, so we should err on the side of caution and choose the surer option (Buchak 2017, pp. 21–24). Or perhaps we have a duty to guard against the worst outcome (Keeney 1980), which is made possible by the probabilistic option – for instance, if the probabilistic rescue fails, it creates the worst possible outcome of all 50 people dying. Secondly, moral risk seeking might be supported by the value of solidarity with others at risk – we might be obliged to save everyone at some significant and proportionate probability, rather than saving some subset with a higher probability (Kamm 1993, pp. 124–126).[3] Thirdly, I'm unaware of explicit arguments for moral risk-neutrality, but we get risk-neutrality if we assume expected utility theory and combine it with the assumption that the

averse. A question then naturally arises: how does the moral analogue of her theory fit in my taxonomy? In my view, this theory would be a second-order one – it says we are rationally permitted to deem it merely permissible to choose a sure option over a probabilistic one, for instance – so it falls outside the scope of my taxonomy, which only aims to categorise first-order theories about the moral status of risky options. I see the relation between this moral analogue of Buchak's theory and the moral risk preferences as akin to the relation between theories of moral uncertainty and first-order moral theories like utilitarianism and Kantianism. I concede too that my empirical results below cannot adjudicate between the moral analogue of Buchak's theory and its competitors (for instance, a theory which says that moral risk neutrality is rationally required). Thanks to an anonymous reviewer for getting me to clarify this.

[2] The size of a benefit/harm is analogous to the amount of money in the non-moral cases, so it should be specified in entirely non-moral terms – such as in terms of the number of people who survive or die in an outcome. This is important so as not to prejudge the question of whether the risk preference is due to our valuing of the outcomes or our weighing of the probabilities. See Buchak (2013, p. 21) for similar justification.

[3] Proportionate probabilities are proportional to the moral weight of each group that could be saved. Moral risk-seeking is sometimes also supported by other rationales – that people should have a non-zero probability of survival, that we are obliged to disperse probability of survival over a larger group rather than concentrate it on a few, that we should equalise ex ante chances as much as possible, or that we should let chance decide peoples' fates, other things being equal (thus favouring options conferring probabilities that are neither 0% nor 100%) (Daniels 2015; Keeney 1980, pp. 529–532; Dreisbach and Guevara 2019, p. 619). These rationales support risk-seeking when we compare the probabilistic option with a different kind of sure option, where which rigidly designated individuals will survive/die has been picked out before the agent's choice (e.g. in the rescue case, if 45 people are stuck in one container, and 5 are stuck in another, and we can try to rescue one or both containers). In such cases, the probabilistic option satisfies more rights to non-zero probability of survival, it disperses the probability of survival over a larger group, and it equalises ex ante chances more, and it lets chance decide the fates of more people. In contrast, when the probabilistic option is compared with a sure option that picks who will live or die by a lottery (as described in the main cases), the probabilistic and risky options satisfy these other rationales equally well.

moral utility is a linear function of the benefit/harm in question (this seems especially plausible when multiple human lives are concerned) (Jackson 1991; Otsuka 2015, pp. 91–92).

Our moral judgments about concrete risky cases (such as the rescue case just presented) can also provide evidence for which moral risk preference to adopt. Case judgments are widely accepted as evidence in ethics, and their empirical study is well-warranted across a wide range of meta-ethical positions (Kahane 2013). I now argue, however, that existing empirical work hasn't studied our judgments about risky cases in a way that sheds light on which moral risk preference might be driving them. First, this requires studying subjects' *moral* judgments about what we should do in risky cases – as distinct from what subjects merely prefer to do[4] (which incorporates moral and non-moral considerations), and from their prudential judgments about how we should choose for others (which concern what's good from their self-interested point of view). This rules out a vast empirical literature which presents risky scenarios to subjects and asks which action they 'favour' or 'prefer', whether as themselves or as social planners (Rheinberger 2010; Kemel and Paraschiv 2018; Abrahamsson and Johansson 2006; Tversky and Kahneman 1981). These elicit mere preferences rather than moral judgments, and can inform the ethics of risk only on the assumption that subjects always prefer to act in accordance with their moral judgments. Of course, this doesn't hold generally: subjects might prefer not to choose an option they think is morally required, if for instance they are squeamish about its possible consequences.[5]

Secondly, the judgments must be elicited in a sufficiently fine-grained way to discriminate between the moral risk preferences – subjects must be given a chance to indicate whether they think an option is required, merely permissible, or forbidden (which is what the difference between the risk preferences consists in). This means that many methods of elicitation are too coarse-grained. If we asked subjects how confident they are that they 'should' pick the probabilistic option (Ryazanov et al. 2021), a high-confidence response could indicate either risk seeking (probabilistic option is required) or risk neutrality (probabilistic option is permitted, but so is the sure one); while a middling-confidence response could either indicate uncertainty between risk-aversion and risk-seeking, or confidence in risk-neutrality. If we asked which option is 'morally better' (Shou and Song 2017), this prevents discrimination between risk-seeking and risk-neutrality, since two options can be merely permissible even when one is better than the other. If we asked subjects to rate the permissibility of an action on a Likert scale from permissible to impermissible (Ryazanov et al. 2018), we don't get any information about whether they also think it's required, failing to measure any potential moral risk-seeking.

For these reasons, I believe there is a gap in the study of our moral risk preferences. To properly elicit them, we must examine subjects' moral judgments (rather than their

[4] It is thus somewhat confusing to speak of moral risk *preferences*. To be clear, I am using the term to refer to the moral analogues of the risk preferences – that is, subjects' moral judgments about what we should do in risky cases.

[5] See Kahane and Shackel (2010), who make this point about a different kind of study.

preferences or prudential intuitions) at a fine-enough level of grain (to distinguish between an option's being forbidden, merely permissible, or required). In sections 3 and 4, I propose a study which does just that.

2. The Reliability of Our Judgments about Risky Cases

Properly elicited case judgments can be a source of evidence in ethics, but their support can also be undermined by evidence of unreliability. In this section, I outline two challenges to the reliability of our judgments about risky cases, one of which I will focus on for my empirical study.

First, our judgments about risky cases could be unreliable because they exhibit framing effects – where the choiceworthiness of an option changes depending on whether it's framed in terms of gains or losses. In the famous Asian Disease Problem, for instance, Tversky and Kahneman find that framing options in terms of gains (number of people saved) versus losses (number who die) changed subjects' preferences over what seemed to be the same policy options (Tversky and Kahneman 1981, p. 453). Philosophers have debated whether these findings support a debunking explanation for our moral judgments about doing and allowing (Dreisbach and Guevara 2019; Horowitz 1998; Kamm 1998; Van Roojen 1999; Sinnott-Armstrong 2007). Perhaps the findings also debunk our moral judgments about risk, since the Asian Disease Problem involves probabilistic options too. Mandel argues, however, that the framing effect observed is due to an unwanted implicature, such that different frames in fact describe different options. He argues that when the fate of some is left unspecified – for instance, when the gains frame says that '400 are saved' out of 600 people – subjects adopt a lower-bound reading of the number, imagining a case where *at least* 400 are saved. If this is right, then subjects adopting a lower-bound reading of the loss frame – reading '200 die' out of 600 people as *at least* 200 dying – would be considering a substantively different option, rather than one that is merely framed differently. In support of this explanation, Mandel finds that framing effects disappear when the fate of everyone in the Asian Disease Problem is specified – that is, when outcomes are described completely as '400 are saved, 200 die', or as '*exactly* 400 will be saved' (Mandel 2014, pp. 1189–1190). Because framing effects could be removed in this way, I set them aside in this paper. (I do however take care to specify outcomes completely, as Mandel has.)

Secondly, our judgments about risky cases might be unreliable because which risk preference we adopt depends in a problematic way on the probabilities involved. To illustrate, consider the following results from a study of preferences over monetary gambles: subjects were on average willing to pay $63 for a gamble that has a 90% probability of paying $100 and 10% probability of paying $0 – that is, they are risk-averse when the probability of winning $100 was high. On the other hand, they were on average willing to pay $10 for a gamble with a 5% probability of paying $100 and 95% probability of paying $0 – they are risk-seeking when the probability of winning $100 was low (Barberis 2013, p. 177; Gonzalez and Wu 1999). One prominent explanation of this is that in the first gamble (90% probability of paying $100), we underweight the 90% probability of success relative to how expected utility theory weighs probabilities, while in the second gamble (5% probability of paying $100), we overweight the 5% probability (Barberis 2013). Generally, it appears that we have risk-averse preferences over gambles with moderate to high probabilities of gains, and risk-seeking preferences over gambles with low probabilities of gains.[6] In the moral domain, Ryazanov et al. (2021) find evidence of probability dependence in our moral judgments about risk, but at a coarser level of grain which doesn't distinguish between the moral risk preferences as I have defined them. My study below builds on theirs, by examining whether and how the moral risk preferences change depending on the probabilities. I also consider whether probability dependence could support novel debunking arguments against our moral judgments about risk – for instance because these judgments exhibit a problematic inconsistency across probability levels. These arguments are best assessed along with the empirical results, so I defer their detailed discussion to the next section.

3. An Empirical Study of our Moral Risk Preferences

I now outline a study which aims to elicit our moral risk preferences and to assess their reliability. I recruited 400 subjects from Amazon Mechanical Turk to give their moral judgments about risky cases. 200 subjects considered a case of risky benefits, while 200 considered a case of risky harm.[7] The risky benefits case is just the rescue case from earlier. To measure any potential probability dependence, subjects who saw this case were randomly assigned one of two conditions: in the '90% probability' condition, the risky rescue has a 90% probability of saving everyone (50 people), as compared with a sure rescue with a 100% probability of saving exactly 45 (the same expected number of lives). In the '10% probability' condition, the risky rescue has a 10% probability of saving 50, and the sure rescue has a 100% probability of saving exactly 5. The 90% probability case reads as follows [10% probability variant in square brackets]:

> You are operating a ship alone in a furious storm, and you receive a call for help from a remote island, where you find 50 people awaiting rescue. You are unsure about how many people your ship can carry safely, and can attempt the following rescue options (assume you do not suffer any costs from choosing either option).
>
> > Probabilistic Rescue with a 90% [10%] probability that 50 people survive (and no one dies), and 10% [90%] probability that no one survives (and 50 die).
> >
> > Sure Rescue with 100% probability that 45 people survive (and 5 die)[that 5 people survive (and 45 die)]. Assume that in this option, you hold a lottery to

[6] See Fehr-Duda and Epper (2011) for an in-depth review of this phenomenon.

[7] Subjects were located in the United States, had more than 90% task approval rating, and had more than 50 approved tasks. They were paid US$0.50 for their participation. This research was preregistered with the Open Science Framework (see https://osf.io/y3vwr for the initial submission, and https://osf.io/txvpb for a further study following reviewer suggestions).

> > pick which 45 will survive and which 5 will die[which 5 will survive and which 45 will die].

Another group of subjects considered a risky harm case – which has the same structure, except it involves what intuitively counts as a harm rather than a benefit to the agents involved. As before, subjects were randomly assigned to a '90% probability' and a 10% probability' condition.[8] The 90% probability condition of the harms case reads as follows [10% probability variant in square brackets]:

> You are a bystander in a factory accident where some toxic gas has been released on a floor with 50 people. You are unsure about how toxic the gas is, and can either disperse the gas to everyone (resulting in a Probabilistic Harm) or concentrate it on some people (resulting in a Sure Harm). That is, you can attempt one of the following options (assume you do not suffer any costs from choosing either option).
>
> > Probabilistic Harm with a 90% [10%] probability that 50 people survive (and no one dies), and 10% [90%] probability that no one survives (and 50 die).
> >
> > Sure Harm with 100% probability that 45 people survive (and 5 die)[that 5 people survive (and 45 die)]. Assume that in this option, you hold a lottery to pick which 45 will survive and which 5 will die [which 5 will survive and which 45 will die].

All subjects were then asked whether they thought they were morally required to choose the probabilistic option, required to choose the sure option, or permitted to choose either. They then also answered some demographic questions, and an attention check question.

3.1. Pooled Results

I received 201 usable responses – 98 for the benefits case and 103 for the harms case – from subjects who passed the attention check.[9] To measure subjects' moral risk preferences, I examine the pattern of responses to the cases, pooled over the probability levels (Fig. 1 and 2). I observe a considerable number of subjects who were risk-seeking and risk neutral, and that moral risk-aversion was the least popular in both kinds of cases. Chi-squared goodness of fit tests were performed to determine whether the proportion of subjects choosing each risk preference was equal. The data did not reject the null hypothesis that subject responses were

[8] For ease of reference, I formulated the cases so that the 90% condition of the risky harms case has the same probabilities and outcomes as the 90% condition of the risky benefits case: both involve a choice between a risky option with 90% probability that exactly 50 survive and 10% probability that exactly no one survives, and a safe option with 100% probability that exactly 45 survive. The same applies with the 10% condition.

[9] As can be seen, a significant proportion of subjects (about 50%) failed the attention check question, reducing the effective sample size for my later analyses and impacting their probative value. I still believe, however, that my results provide some provisional evidence concerning the ethics of risk.

equally distributed across the different risk preferences, in both the benefit (χ^2(2, N = 98) = 5.47, p = .0649) and harm (χ^2(2, N = 103) = 5.03, p = .0809) cases. The low p-values (probability of obtaining this data, given that subjects were in fact equally distributed across risk preferences) and the distributions (in Fig. 1 and 2) indicate that my data came close to rejecting this null hypothesis because subjects were skewed towards risk-seeking and risk-neutrality. Still, even if we took the data to only indicate an equal distribution across different moral risk preferences, this suggests significant heterogeneity in subjects' moral risk preferences – in particular, that moral risk-seeking and risk-neutrality are quite popular. This constitutes *pro tanto* evidence in support of theories prescribing these risk preferences, and surprising evidence against moral risk-aversion, which has been favoured by many philosophers.

Fig. 1 Pooled Benefit Case Responses

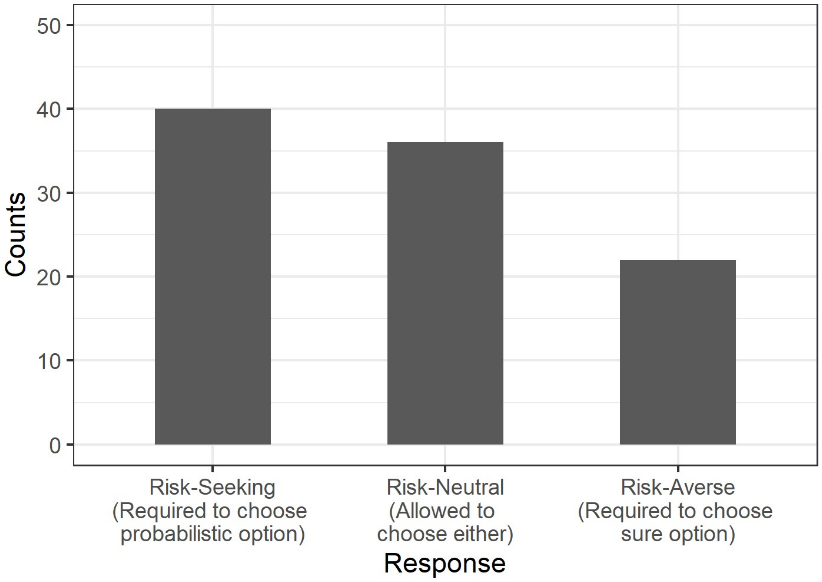
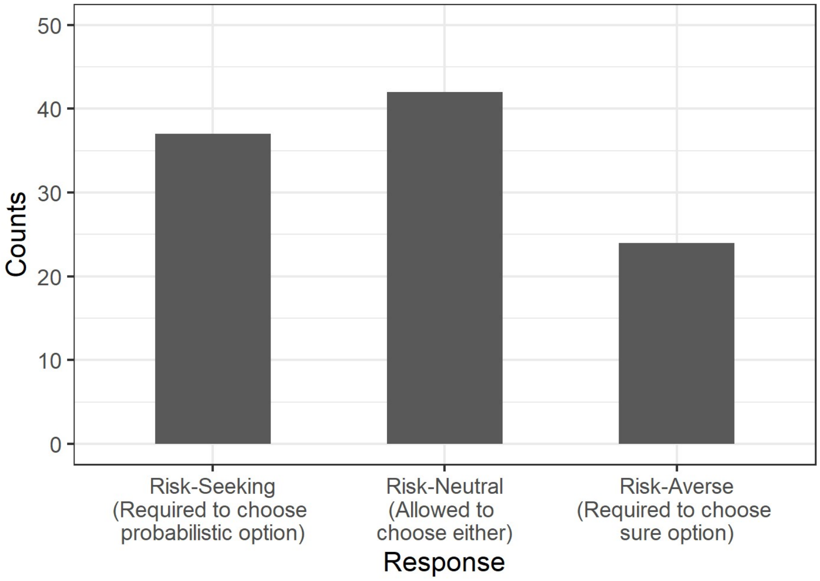


Fig. 2 Pooled Harm Case Responses



One could try to undermine the reliability of these responses on the basis of the risk preference they exhibit. Consider, for instance, a proponent of risk-aversion who argues: probabilistic options should never be morally permitted (or required) over sure options that have the same expected effect. Our judgments about risk do sometimes permit or require choosing the probabilistic option. Therefore, our judgments are unreliable. This kind of debunking argument is problematic because it lacks evidential value in the relevant context. The reason why we want to assess the reliability of our moral judgments about risk is to find out the truth about how we should weigh these risks. So this assessment occurs in an evidential context where it remains open how we should weigh such risks. This debunking argument prejudges this very issue, however, since it assumes from the outset that we should be morally risk-averse. To put the point differently, this argument wouldn't convince someone who is an agnostic about which risk preference we should adopt, since an agnostic wouldn't have been independently convinced of moral risk-aversion in the first place. Things might be different, however, if there were independent evidence – evidence that does not rely on our case judgments – in favour of moral risk-aversion. But if we're admitting independent arguments in favour of risk-aversion, we would also need to take into account independent arguments for other risk preferences too: for instance, we might argue from diachronic consistency to expected utility theory (Hammond 1988; Quiggin 1993, pp.121–124), and then combine that with the assumption that the moral value should be a linear function of the number of lives at stake (which we might have an independent argument for), to get the moral risk-neutrality with respect to lives.

3.2. Probability Dependence and Debunking

I turn now to whether our moral risk preferences exhibit a problematic dependence on the probabilistic option's absolute probability of success. Recall how this dependence manifests in monetary gambles: subjects were risk-seeking at low probabilities of gains (they preferred the 5% probability of winning $100 over getting its expected return for sure), and risk-averse at high probabilities of gains (rather than choosing the 90% probability of winning $100, subjects preferred getting its expected return for sure).

If we observed the same kind of probability dependence in our moral judgments about risk, then we might have independent grounds to debunk them. This is because many of the rationales for the moral risk preferences (as canvassed in section 1) justify a uniform risk preference across different absolute probability levels. For instance, if we should err on the side of caution and choose the sure option, this applies just as well in the choice between a 90% probability option and a sure option which saves the same expected number, as it does in the choice between a 10% probability option and a sure option saving the same expected number. That is, the justification for moral risk-aversion applies equally well in the 90% and 10% probability conditions. Similarly, if we have a duty to guard against the worst outcome, this also justifies uniform moral risk-aversion across both the 90% and the 10% probability conditions; if we assume expected utility theory and argue that moral utility is a linear function of lives, this justifies uniform moral risk-neutrality across different probabilities. The disjunction of these rationales thus supports a debunking argument against probability-dependent moral judgments about risky cases. This debunking argument relies on weaker assumptions – it only assumes that risk preference should be consistent across probability levels, but not what the correct risk preference should be at any level – and so is less likely to be evidentially redundant.

There are, however, also rationales that justify probability-dependent moral risk preferences. Consider the rationale for moral risk-seeking, which says solidarity dictates that we should save everyone at some significant and proportionate probability. Perhaps 90% would count as a significant probability, whereas 10% would not – in which case this rationale only justifies risk-seeking in the 90% probability condition, but not in the 10% condition. Notice, however, that it justifies a pattern of dependence that's the reverse of what we see in monetary gambles – it justifies risk-seeking at *high* probabilities of success, rather than at low ones.

To measure whether (and how) our moral risk preferences depend on the probabilities, I look at subjects' responses disaggregated over the 90% and 10% probability conditions. I then check to see if the pattern of subjects' responses changed depending on the absolute probability of success of the probabilistic option. I find no evidence that the pattern of moral risk preferences depends on absolute probability level, in either the benefit or the harm cases (see . 3 and 4). For each type of case, I conducted a chi-squared test of independence, and found no statistically significant relationship between risk preference and absolute probability of success of the probabilistic option ($\chi^2(2, N = 98) = 0.918$, $p = .632$ for the benefits case, and $\chi^2(2, N = 103) = 1.50$, $p = .473$ for the harms case).

Fig. 3 Benefit Case Responses by Probability

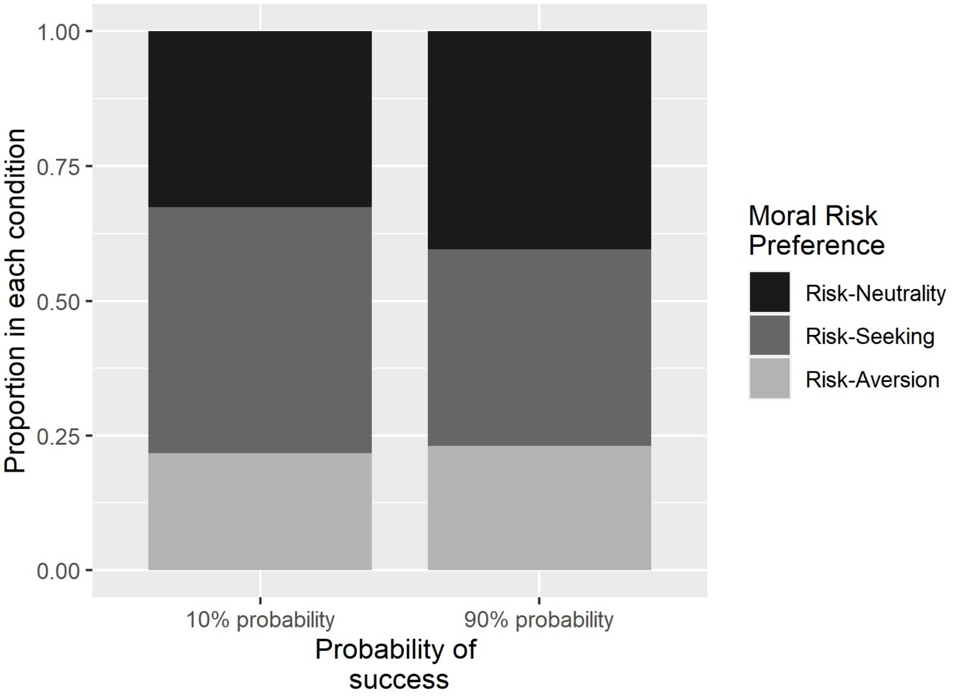
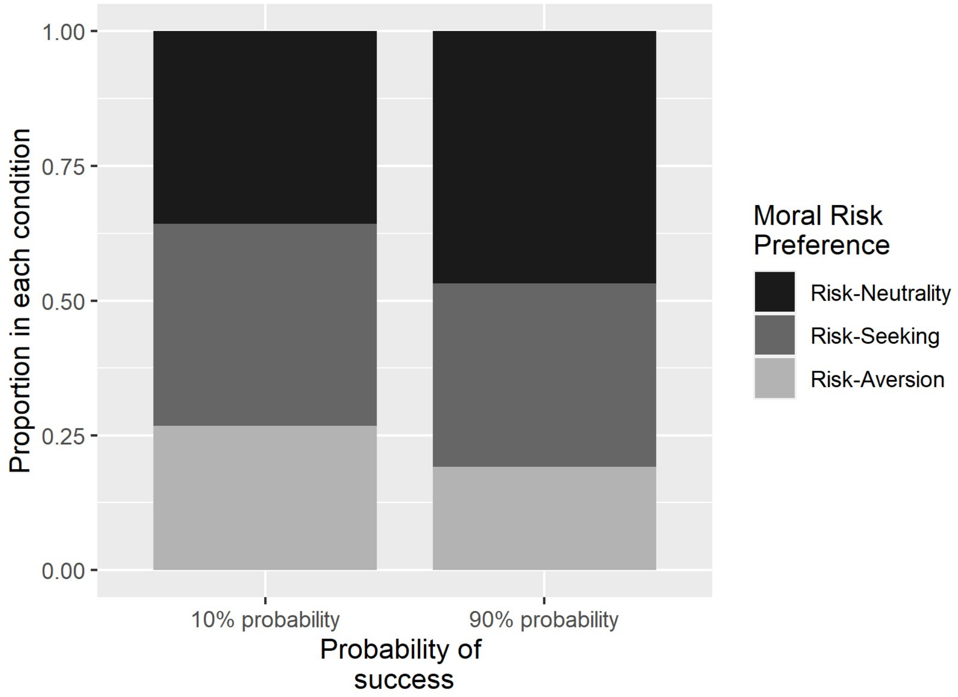


Fig. 4 Harm Case Responses by Probability



My results thus provide an empirical defence against the debunking argument from probability dependence. If we did find evidence of probability dependence in our moral judgments (and we assumed the disjunction of rationales supporting a uniform risk preference), then we can conclude that these judgments are unreliable. In my study, I didn't find evidence of such probability dependence, so this debunking argument fails.

However this defence is only limited to the cases I've studied, leaving open the possibility that judgments about other cases could still be debunked. For instance, Ryazanov et al. (2021) find probability dependence in judgments about cases where an agent can impose a risky harm on some people in order to save others with certainty. Our studies are not directly comparable, because their cases involve options that impose a mix of benefits and harms, whereas my cases involve options with only benefits or only harms; they also measure subjects' moral risk preferences using the average confidence (across the population) that a certain option should be taken, and do not distinguish between moral requirement and permissibility like I have. More research is needed to determine whether these features are responsible for our differing results, and whether their results can support a debunking argument from probability dependence too. The scope of any debunking conclusion is thus importantly constrained by the empirical evidence. Are all our moral judgments about risk debunked, regardless of whether they concern benefits or harms? Or is it only judgments about imposing risky harms specifically? Or, more narrowly still, our judgments about imposing risky harms in order to save some other agents? I believe the empirical evidence will determine the level of grain at which debunking occurs (if debunking does occur at all).[10] So perhaps my conclusions won't carry over to other as-yet-unstudied judgments, for instance about cases involving much smaller probabilities – much smaller than 10%, say 0.001% instead – of very large outcomes, where moral risk-neutrality seems most controversial. But I take my investigation to have at least shed light on more ordinary cases of risk, with probabilities as low as 10%.

Furthermore, even if a debunking argument from probability dependence succeeds, there is room to correct the relevant moral judgments.[11] Suppose we were confident that there must be a uniform moral risk preference across different probabilities, but we found that subjects adopted different risk preferences at different probability levels (so that at least one of their judgments is mistaken). If we're also more confident in judgments about one of these probabilities than the other, then we can generalize the risk preference exhibited in the probability level that we're more confident about. For instance, if we're more confident about judgments made in the 10% probability condition, then we should disregard judgments from the 90% probability condition, and adopt the moral risk preference expressed in the 10% probability condition throughout. Alternatively, if we were equally confident in judgments from both these conditions, then we should suspend judgment for now and look for other evidence – evidence that's independent of our case judgments – concerning the ethics of risk.

Finally, my findings also bear on secondary research questions about the processes underlying our moral judgments about risky cases. The results here indicate that our moral risk preferences behave quite differently from our non-moral risk preferences over monetary gambles – since the latter exhibit probability dependence, whereas I found evidence that the former do not. Thus the overweighting and underweighting explanation given for our non-moral risk preferences (as outlined in section 2) cannot be the whole story – there are likely different, or additional, processes operating to produce our moral risk preferences. More research needs to be done to understand when and how these processes work in the moral domain.

4. A Further Study of Moral Risk Preferences over Intrapersonal Tradeoffs

The central cases I've studied so far are risky interpersonal tradeoffs – where we have to choose between gambles where the interests of multiple people conflict. But our moral risk preferences also apply to risky intrapersonal tradeoffs, where only the interests of a single person are at stake. We might wonder whether the same pattern of risk preferences obtains for intrapersonal tradeoffs – or, say, if moral risk-aversion is more prevalent there – and whether the debunking argument from probability dependence has any purchase on our judgments about intrapersonal cases.[12] To answer these questions, I conducted a further study of our moral risk preferences over intrapersonal tradeoffs, using the following analogous

[10] See Machery (2017, pp. 97–99) and Liao et al. (2012, pp. 667–668) for related discussion.

[11] Vavova (2021) raises the possibility of correction in the context of evolutionary debunking.

[12] Some prominent cases in the ethics of risk involve verdicts of risk-aversion in intrapersonal tradeoffs – for instance see Buchak (2017, pp. 630–633). Thanks here to an anonymous reviewer.

benefit and harm cases. As before, the 90% probability variants are presented [10% variants in square brackets]:

> *Intrapersonal benefit case*
>
> You are a bystander who finds a person unconscious by the road. They were given a poison that will kill them unless counteracted immediately. You can choose from one of two antidotes to give them, as follows (assume that the person is unlikely to die of other causes in the extra years made possible by these options, and assume that you do not suffer any costs from choosing either option).
>
> > Probabilistic Antidote with a 90% [10%] probability that the person will live for another 10 years exactly, and a 10% [90%] probability that the person dies immediately because the antidote has no effect.
> >
> > Sure Antidote with 100% probability that the person will live for another 9 years exactly.[that the person will live for another 1 year exactly.]

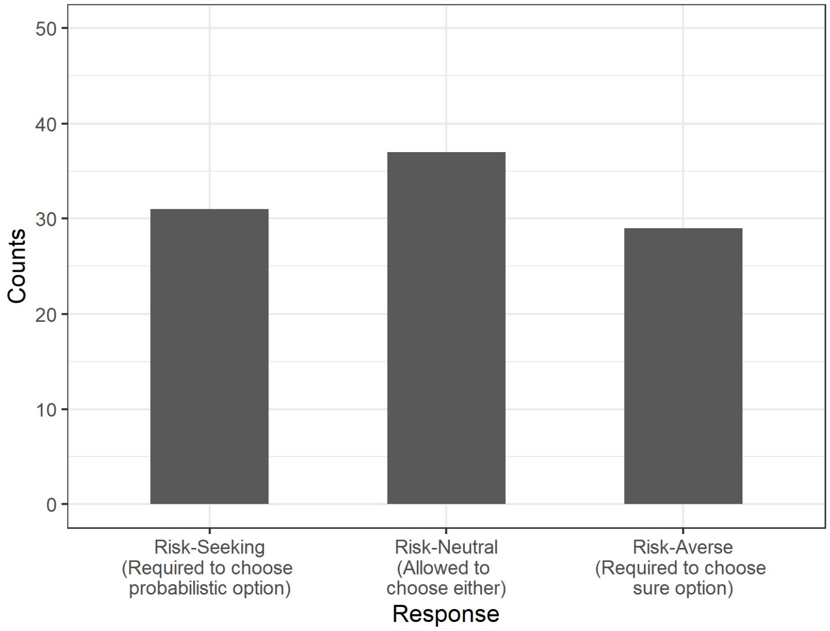
> *Intrapersonal harm case*
>
> You are a bystander in a factory accident where some toxic gas is about to be released into an office with one person. You cannot stop the release of the gas, but you can convert it into one of two forms, which will result in either a Probabilistic Harm or a Sure Harm, as follows (assume that the person is unlikely to die of other causes in the extra years made possible by these options, and assume that you do not suffer any costs from choosing either option).
>
> > Probabilistic Harm with a 90% [10%] probability that the person will live for another 10 years exactly, and a 10% [90%] probability that the person dies immediately.
> >
> > Sure Harm with 100% probability that the person will live for another 9 years exactly.[that the person will live for another 1 year exactly.]

4.1 Pooled Results

For these cases, I recruited another 400 subjects from Mechanical Turk – with 200 seeing a case of intrapersonal benefit tradeoffs, and 200 seeing a case of intrapersonal harm tradeoffs. I received 189 usable responses from subjects who passed the attention check – 97 for the benefits case and 92 for the harms case.

I start by examining subjects' moral risk preferences in these intrapersonal cases, pooled over different probability levels. In the intrapersonal benefit case (Fig. 5), subjects were quite evenly split across the different risk preferences – a chi-squared goodness of fit test did not reject the null hypothesis that subject responses were equally distributed across different risk preferences ($\chi^2(2, N = 97) = 1.07, p = .585$). Notably, the p-value (the probability of obtaining the data observed, given that subjects were in fact equally distributed) is much higher for intrapersonal benefit tradeoffs than for the interpersonal cases or the intrapersonal harms case.[13] This indicates that of all the cases studied, subjects are most likely to be evenly distributed across different risk preferences in the intrapersonal benefits case as compared to all other cases. Moral risk-aversion is thus more popular in intrapersonal benefit tradeoffs, but there also remain significant numbers who are morally risk-seeking or risk-neutral in such cases.

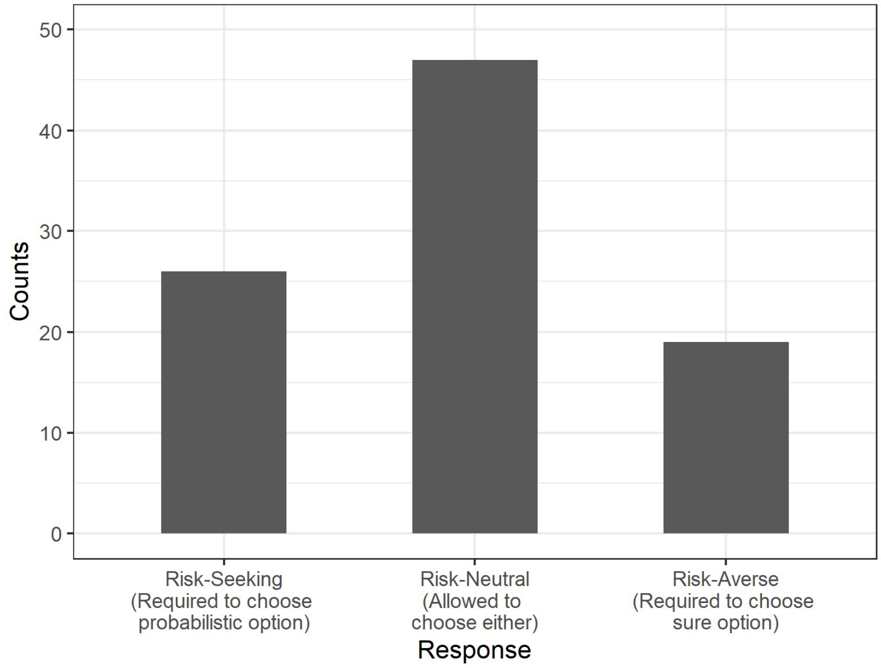
Fig. 5 Pooled Intrapersonal Benefit Case Responses



In contrast, in the intrapersonal harms case, a large and significant proportion of subjects were morally risk-neutral, with smaller proportions being risk-seeking or risk-averse (Fig. 6). A chi-squared goodness of fit test rejected the null hypothesis that subject responses were equally distributed across risk preferences ($\chi^2(2, N = 92) = 13.8, p = .000984$). Interestingly, then, moral risk-aversion is only significantly popular in intrapersonal benefit tradeoffs, but not in intrapersonal harm tradeoffs.

[13] Recall that the p-value for the corresponding chi-squared test for the interpersonal benefits case is 0.0649, and the one for interpersonal harms is 0.0809.

Fig. 6 Pooled Intrapersonal Harm Case Responses



We can summarise my pooled results for intrapersonal and interpersonal cases in the following table (recall: the null hypothesis here is that subjects were equally distributed across different moral risk preferences for that category).

	Intrapersonal	**Interpersonal**
Benefits	Did not reject null hypothesis of equal distribution, p = 0.585	Did not reject null hypothesis of equal distribution, but observed skew towards risk-neutrality and risk-seeking, p = 0.0649
Harms	Rejected null hypothesis of equal distribution, skew towards moral risk-neutrality, p = 0.000984	Did not reject null hypothesis of equal distribution, but observed skew towards risk-neutrality and risk-seeking, p = 0.0809

The most striking feature of my findings is the considerable popularity of case judgments reflecting moral risk-neutrality and risk-seeking – whether in intrapersonal or interpersonal cases, whether over benefits or harms. In the absence of a plausible debunking argument against these judgments, philosophers would do well to take moral risk-seeking and risk-neutrality seriously in the ethics of risk.

We can also compare the pattern of moral risk preferences between intrapersonal and interpersonal tradeoffs (that is, comparing across rows in the above table). Start with the comparison within the category of benefits: in both intrapersonal and interpersonal benefit tradeoffs, my data did not reject the null hypothesis that subjects were equally distributed across different risk preferences. Nonetheless, a comparison of p-values indicates that intrapersonal benefit tradeoffs were much more likely to elicit an equal distribution of moral risk preferences – in particular, more subjects endorsing moral risk-aversion – than interpersonal benefit tradeoffs (where I observed a statistically insignificant skew towards risk-seeking and risk-neutrality). Next consider the comparison within the category of harms: when trading off harms to a single person, a large and significant majority of subjects were morally risk-neutral, whereas when trading off harms to multiple people, subjects were more likely equally distributed across different risk preferences (though I also observed a statistically insignificant skew towards risk-seeking and risk-neutrality).

We might be able to reconcile and rationalize some of these differences. For instance, we could adopt a consequentialist theory that assigns extra utility to the chance of saving everyone (to account for risk-seeking or risk-neutrality in interpersonal tradeoffs) but that also uses a risk-averse decision theory to weigh utilities (to account for risk-aversion in intrapersonal tradeoffs); or we could assign diminishing marginal moral value to outcomes in intrapersonal – but not interpersonal – tradeoffs.[14]

4.2 Probability Dependence in Intrapersonal Tradeoffs

I now investigate the potential probability dependence of risk preferences in intrapersonal cases, by looking at subject responses disaggregated over the 90% and 10% probability conditions. I also find no evidence that the pattern of moral risk preferences over intrapersonal tradeoffs depends on absolute probability level, in the benefit or harm cases (see Fig. 7 and 8). For each case, a chi-squared test of independence found no statistically significant relationship between risk preference and absolute probability of success of the probabilistic option ($\chi^2(2, N = 97) = 0.961$, $p = .618$ for intrapersonal benefits, $\chi^2(2, N = 92) = 2.25$, $p = .325$ for intrapersonal harms.)

[14] Thanks here to an anonymous reviewer.

Fig. 7 Intrapersonal Benefit Case Responses by Probability

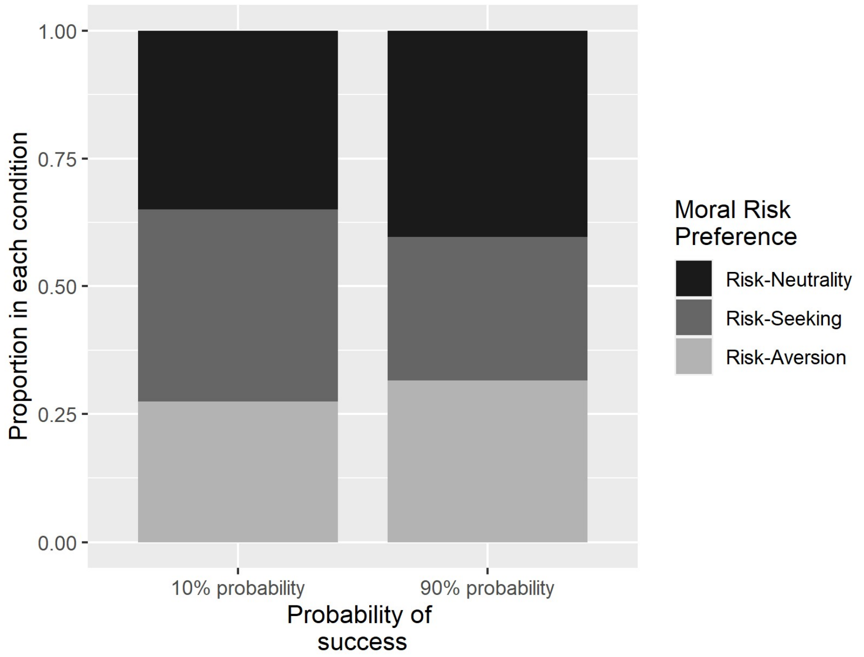
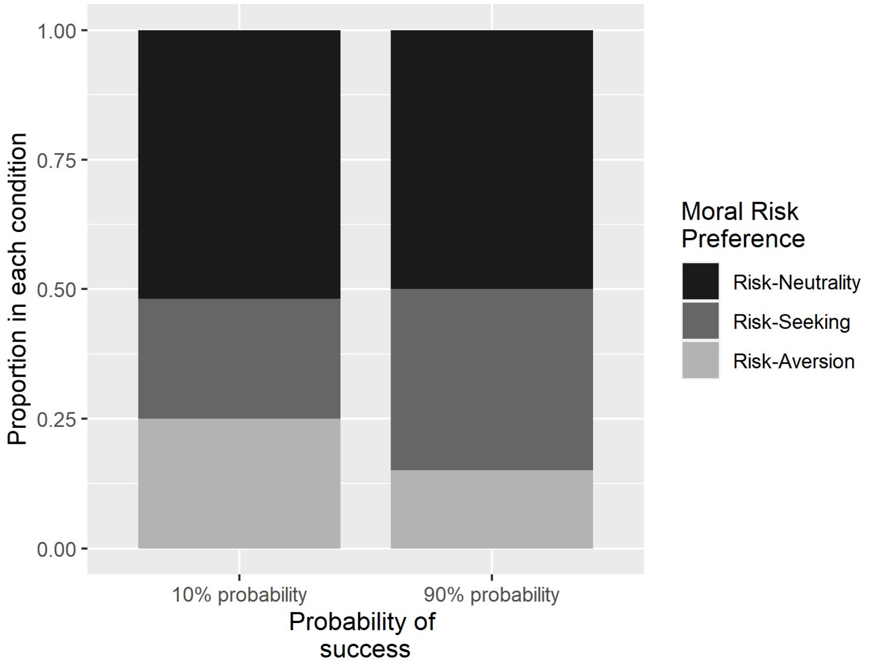


Fig. 8 Intrapersonal Harm Case Responses by Probability



As before, then, my results support an empirical defence of our moral risk preferences against a debunking argument from probability dependence. I found no evidence of problematic probability dependence in our judgments about risky intrapersonal tradeoffs – though, again, this does not preclude the possibility that some other judgments about risk are debunked on the same grounds.

5. Conclusion

In this paper, I cleared the ground for an empirical study of our moral risk preferences, and then detailed the method and results of my study. I started by defining the moral risk preferences in terms of moral requirement and permissibility – I then argued that the empirical literature hasn't adequately studied these moral risk preferences, so defined. In eliciting these risk preferences using concrete cases, I found moral risk-seeking and risk-neutrality to be surprisingly popular in both interpersonal and intrapersonal risky cases. This offers defeasible evidence in support of theories that prescribe moral risk-seeking and risk-neutrality. I argued too that we cannot discount subjects' risk preferences merely on the basis of their content – for instance, we cannot discredit their risk-neutrality on grounds that moral risk-aversion is correct – because this fails in the relevant evidential context, where we are agnostic about how we should weigh the risks.

Instead, I looked for evidence of probability dependence – where our risk preferences depend on the absolute probability of success of the probabilistic option – which could ground an independent debunking argument. I find no evidence indicating that subjects' judgments about cases – whether interpersonal or intrapersonal, whether about benefits or harms – are probability dependent at all. This constitutes an empirical defence against this line of debunking, since I didn't obtain the relevant evidence that would support the debunking argument. This illustrates how, ultimately, the empirical evidence determines whether a debunking conclusion obtains – and if so, what the scope of such a conclusion would be. Regardless of whether we're ultimately hopeful or pessimistic about the reliability of our moral risk preferences, I hope for this paper to have provided a case study in how the empirical evidence can usefully supplement armchair theorising about the ethics of risk.

National University of Singapore and
Monash University
3 Arts Link, Block AS3
Singapore 117570
shanglong@nus.edu.sg

NOTES

Many thanks to three anonymous reviewers for the *American Philosophical Quarterly* for their perceptive and detailed comments. I've also benefitted greatly from feedback from audiences at the 2019 edition of the ANU-Humboldt-Princeton Summer Institute and at the 2023 ANU PPE workshop. I would like to thank these audiences, and thanks especially to Christian Barry, Seth Lazar, Alan Hájek, Katie Steele, Toby Handfield, Ramon Das, James Willoughby, Kirsten Mann, Devon Cass, Chris Lernpass, Oliver Rawle, Jeremy Strasser, Sean Donahue, Zach Barnett, Jesse Hambly, Hezki Symonds, and Josef Holden for discussions of this paper in its various forms (including an early non-empirical version). This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

REFERENCES

Abrahamsson, Marcus, and Henrik Johansson. 2006. "Risk Preferences Regarding Multiple Fatalities and Some Implications for Societal Risk Decision Making—An Empirical Study," *Journal of Risk Research,* vol. 9, no. 7, pp. 703–715.

Barberis, Nicholas. 2013. "Thirty Years of Prospect Theory in Economics: A Review and Assessment," *Journal of Economic Perspectives,* vol. 27, no. 1, pp. 173–196.

Buchak, Lara. 2013. *Risk and Rationality.* (Oxford: Oxford University Press).

———. 2017. "Taking Risks Behind the Veil of Ignorance," *Ethics,* vol. 127, no. 3, pp. 610–644.

Charness, Gary, Uri Gneezy, and Alex Imas. 2013. "Experimental Methods: Eliciting Risk Preferences," *Journal of Economic Behavior & Organization* vol. 87, March, pp. 43–51.

Daniels, Norman. 2015. "Can There Be Moral Force to Favoring an Identified over a Statistical Life?" in *Identified versus Statistical Lives: An Interdisciplinary Perspective*, ed. I. Glenn Cohen, Norman Daniels, and Nir Eyal. (Oxford, UK: Oxford University Press), pp. 110–123.

Dreisbach, Sandra, and Daniel Guevara. 2019. "The Asian Disease Problem and the Ethical Implications Of Prospect Theory," *Noûs,* vol. 53, no. 3, pp. 613–638.

Fehr-Duda, Helga, and Thomas Epper. 2011. "Probability and Risk: Foundations and Economic Implications of Probability-Dependent Risk Preferences," *Annual Review of Economics,* vol. 4, no. 1, pp. 567–593.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function," *Cognitive Psychology*, vol. 38, no. 1, pp. 129–166.

Hammond, Peter J. 1988. "Consequentialist Foundations for Expected Utility," *Theory and Decision,* vol. 25, no. 1, pp. 25–78.

Horowitz, Tamara. 1998. "Philosophical Intuitions and Psychological Theory," *Ethics,* vol. 108 no. 2, pp. 367–85.

Jackson, Frank. 1991. "Decision-Theoretic Consequentialism and the Nearest and Dearest Objection," *Ethics,* vol. 101, no. 3, pp. 461–482.

Kahane, Guy. 2013. "The Armchair and the Trolley: An Argument for Experimental Ethics," *Philosophical Studies,* vol. 162, no. 2, pp. 421–445.

Kahane, Guy, and Nicholas Shackel. 2010. "Methodological Issues in the Neuroscience of Moral Judgement," *Mind & Language,* vol. 25, no. 5, pp. 561–582.

Kamm, F. M. 1993. *Morality, Mortality: Volume 1: Death and Whom to Save It From*. Oxford: Oxford University Press.

———. 1998. "Moral Intuitions, Cognitive Psychology, and the Harming-Versus-Not-Aiding Distinction," *Ethics,* vol. 108, no. 3, pp. 463–488.

Keeney, Ralph L. 1980. "Equity and Public Risk," *Operations Research,* vol. 28, no. 3, pp. 527–34.

Kemel, Emmanuel, and Corina Paraschiv. 2018. "Deciding about Human Lives: An Experimental Measure of Risk Attitudes under Prospect Theory," *Social Choice and Welfare,* vol. 51, no. 1, pp. 163–192.

Liao, S. Matthew, Alex Wiegmann, Joshua Alexander, and Gerard Vong. 2012. "Putting the Trolley in Order: Experimental Philosophy and the Loop Case," *Philosophical Psychology,* vol. 25, no. 5, pp. 661–671.

Machery, Edouard. 2017. *Philosophy Within Its Proper Bounds*. Oxford, UK: Oxford University Press.

Mandel, David R. 2014. "Do Framing Effects Reveal Irrational Choice?" *Journal of Experimental Psychology. General,* vol. 143, no. 3, pp. 1185–1198.

Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. 2018. "Risk Preference: A View from Psychology," *Journal of Economic Perspectives,* vol. 32, no. 2, pp. 155–172.

Otsuka, Michael. 2015. "Risking Life and Limb: How to Discount Harms by Their Improbability." in *Identified versus Statistical Lives: An Interdisciplinary Perspective*, ed. I. Glenn Cohen, Norman Daniels, and Nir Eyal. (Oxford, UK: Oxford University Press), pp. 77–93.

Quiggin, John. 1993. *Generalized Expected Utility Theory: The Rank-Dependent Model*. Springer Netherlands.

Rheinberger, Christoph M. 2010. "Experimental Evidence Against the Paradigm of Mortality Risk Aversion," *Risk Analysis,* vol. 30, no. 4, pp. 590–604.

Ryazanov, Arseny A., Jonathan Knutzen, Samuel C. Rickless, Nicholas J. S. Christenfeld, and Dana Kay Nelkin. 2018. "Intuitive Probabilities and the Limitation of Moral Imagination," *Cognitive Science,* vol. 42, no. S1, pp. 38–68.

Ryazanov, Arseny A., Shawn Tinghao Wang, Samuel C. Rickless, Craig R. M. McKenzie, and Dana Kay Nelkin. 2021. "Sensitivity to Shifts in Probability of Harm and Benefit in Moral Dilemmas," *Cognition,* vol. 209, April, p. 104548.

Shou, Yiyun, and Fei Song. 2017. "Decisions in Moral Dilemmas: The Influence of Subjective Beliefs in Outcome Probabilities," *Judgment and Decision Making,* vol. 12, no. 5, pp. 481–490.

Sinnott-Armstrong, Walter. 2007. “Framing Moral Intuitions.” In *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity,* ed. Walter Sinnott-Armstrong. (Cambridge: MIT Press) pp. 47–76.

Tversky, Amos, and Daniel Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice,” *Science,* vol. 211, no. 4481, pp. 453–458.

Van Roojen, Mark. 1999. “Reflective Moral Equilibrium and Psychological Theory,” *Ethics,* vol. 109, no. 4, pp. 846–857.

Vavova, Katia. 2021. “The Limits of Rational Belief Revision: A Dilemma for the Darwinian Debunker,” *Noûs,* vol. 55, no. 3, pp. 717–734.

Wakker, Peter P. 2010. *Prospect Theory: For Risk and Ambiguity*. Cambridge, UK: Cambridge University Press.