RUNNING HEAD: Philosophers' Biased Judgments Persist

Philosophers' Biased Judgments Persist Despite Training, Expertise and Reflection

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside

Fiery Cushman
Department of Psychology
Harvard University

Authors Note

Eric Schwitzgebel, Department of Philosophy, University of California at Riverside

Fiery Cushman, Department of Psychology, Harvard University

Correspondence concerning this article should be addressed to Eric Schwitzgebel, Department of Philosophy, University of California at Riverside, Riverside, California, 92521-0201, USA. Email: eschwitz at domain- ucr.edu.

Word count (excluding title, abstract, tables, figures, and bibliography): 6675 words

Philosophers' Biased Judgments Persist Despite Training, Expertise and Reflection

Abstract

We examined the effects of framing and order of presentation on professional philosophers' judgments about a moral puzzle case (the "trolley problem") and a version of the Tversky & Kahneman "Asian disease" scenario. Professional philosophers exhibited substantial framing effects and order effects, and were no less subject to such effects than was a comparison group of non-philosopher academic participants. Framing and order effects were not reduced by a forced delay during which participants were encouraged to consider "different variants of the scenario or different ways of describing the case". Nor were framing and order effects lower among participants reporting familiarity with the trolley problem or with loss-aversion framing effects, nor among those reporting having had a stable opinion on the issues before participating the experiment, nor among those reporting expertise on the very issues in question. Thus, for these scenario types, neither framing effects nor order effects appear to be reduced even by high levels of academic expertise.

Keywords: Doctrine of the Double Effect, experimental philosophy, expertise, framing effects, loss aversion, morality, order effects, reasoning, social cognition

Philosophers' Biased Judgments Persist Despite Training, Expertise and Reflection

Schwitzgebel and Cushman (2012) report that professional philosophers are no less subject to order effects on their judgments about familiar types of moral dilemmas (such as the famous "trolley problem") than are non-philosophers: When scenario pairs were presented in order AB, participants responded differently than when the same scenario pairs were presented in order BA, and the philosophers showed no less of a shift than did the comparison groups, across several types of scenario. As suggested by Sinnott-Armstrong (2008), Weinberg, Gonnerman, Buckner, and Alexander (2010), Liao, Wiegmann, Alexander, and Vong (2012), Schwitzgebel and Cushman (2012), Tobia, Buckwalter, and Stich (2013), and Mizrahi (2015), if philosophers' judgments about puzzle cases in their area of expertise are highly influenced by presumably irrelevant factors such as order of presentation or superficial differences in phrasing, that creates a prima facie challenge to certain optimistic views about philosophical expertise in assessing such scenarios – views of the sort expressed in Ludwig (2007), Grundmann (2010), and Williamson (2011; though see Buckwalter, forthcoming; Nado, forthcoming). It would also suggest a striking persistence of biased decision-making despite extensive training both in logical reasoning in general and in closely related task types in particular.

In the present study we attempt to establish boundary conditions on this effect. Specifically, we attempted to replicate our original effect, but then to reduce its magnitude in four ways: by (a) limiting the target group to philosophers with *expertise specifically on the types of dilemma in question;* or (b) by limiting the target group to philosophers who report *having stable opinions on the matter* (see discussion in Wright, 2010, 2013; Rini, 2015); or (c) by encouraging participants to give *reflective responses, and enforcing a delay for reflection before*

response; or (d) by presenting pairs of *scenarios that differ primarily in phrasing* rather than in the relevant content of the scenario. To the extent the magnitude of the order effect is reduced by any of factors (a)-(d), that might encourage optimism about expert philosophical judgment appropriately restricted. Conversely, to the extent the magnitude of the order effect is not so reduced, that deepens the skeptical challenge.

Beyond its application to philosophical methods, our study of philosophical decision-making has a broader application to cognitive science. Over the past decades researchers have extensively documented the role of heuristics and biases in human judgment and decision-making. Often they have also argued that we would be better off if we could effectively substitute unbiased procedures (Baron, 2000; Thaler & Sunstein, 2008; Kahneman, 2011; Greene 2014). Fewer studies address how this might be accomplished, especially in complex domains without clear feedback procedures. Here, we test some likely possibilities: Slow people down, have them think reflectively and counterfactually; familiarize them with the specific types of decisions in question; provide them extensive instruction and practice in general logical reasoning. Which, if any, of these approaches reliably reduce cognitive bias?

Prior research

Our previous study yielded two main findings. First, and receiving the most straightforward empirical support, we found that professional academic philosophers' and academic non-philosophers' moral judgments were similarly influenced by order of presentation. We tested three categories of moral judgments: several versions of the trolley problem (e.g., the footbridge and switch variants; Foot, 1967; Thomson, 1985; McIntyre, 2004/2011), cases involving moral luck (e.g., degree of blameworthiness when identical conduct such as drunk

driving is either harmless or fatal; Nagel, 1979; Williams, 1981; Nelkin, 2004/2013), and cases that contrast active harm and passive harm (e.g., snatching a life preserver away from a drowning person versus failing to offer that person your own life preserver; Quinn, 1989; Bennett, 1998; Howard-Snyder, 2002/2011). Aggregating across all three types of case we found no evidence that order effects were weaker for philosophers. Moreover, one case in a matched pair was typically more influenced by order than another. For instance, judgments of the switch version of the trolley problem were more strongly influenced by order than judgments of the footbridge version. Consequently, order had an effect on the likelihood that pairs of cases were judged to be *morally equivalent*. For instance, the switch and footbridge cases were more likely to be judged equivalently when presented in the footbridge/switch order than when presented in the switch/footbridge order.

Our second finding concerned the relationship between the judgment of specific vignettes (e.g., the switch and footbridge variants of the trolley problem) and the endorsement of abstract moral principles (e.g., the Doctrine of Double Effect, which purports to justify discrepant judgments between these cases). We hypothesized that participants – both philosophers and non-philosophers – would tend to endorse moral principles in a manner that matches their patterns of judgment. Because order of presentation influenced the likelihood of the cases being judged equivalently, this influence might carry over to influence participants' endorsement of moral principles. For philosophers, we found such an effect for the Doctrine of Double Effect and for a principle asserting the non-equivalency of moral luck cases, but not for a principle asserting the non-equivalency of action/omission cases. For non-philosophers we found precisely the opposite pattern of effects. Moreover, we identified several non-predicted effects of vignette order on endorsement among philosophers (e.g., the order of presentation of moral luck cases affected the

endorsement of the Doctrine of the Double Effect). Overall, these results provided tentative evidence for an effect of order-of-judgment on the endorsement of abstract moral principles, but also suggested that such effects are highly contextually dependent.

Two other empirical studies have explored the relationship between philosophical expertise and bias in moral judgment. Tobia, Buckwalter, and Stich (2013) found that professional philosophers considering moral scenarios were subject to actor-observer biases of about the same magnitude as non-philosophers' (though the groups' biases went in different directions). Tobia, Chapman, and Stich (2013) replicated this result and also found philosophers influenced about as much as were non-philosophers by the presence of a "clean" Lysol odor (though again in different directions). Relatedly, Schulz, Cokely & Feltz (2011) find personality-related differences in philosophical experts' judgments about free will, and Machery (2012) finds subfield-related differences in judgments about linguistic reference.

There is also some research that focuses on the broader question of how expertise affects susceptibility to judgment and decision biases. Reyna, Chick, Corin and Hsia (2014) find that intelligence analysts are, in fact, *more* likely than college students and non-expert adults to exhibit framing effects in the Asian disease problem, and a more comprehensive meta-analysis reveals no significant effects of participant group on the magnitude of framing effects (Kühlberger, 1998). There is also a substantial literature on the fairly limited effects of education on other reasoning tasks, such as the conjunction fallacy and the Wason selection task (Tverksy & Kahneman, 1983; Cheng, Holyoak, Nisbett & Oliver, 1986; Lehman, Lempert & Nisbett, 1988; Ritchhart & Perkins, 2005; Heijltjes, van Gog, Leppink & Paas, 2014). On the other hand, some evidence suggests that philosophers in particular might be unusually skilled at reasoning. Livengood, Sytsma, Feltz, Scheines & Machery (2010) found that philosophers

exhibited superior performance on the Cognitive Reflection Test, a series of simple math problems prone to incorrect intuitive responding (Frederick, 2005), and Kuhn (1991) found that philosophy graduate students were substantially more skilled in evaluating arguments and evidence than were comparison groups, including schoolteachers (though Cheng et al. (1986) find no improvement in Wason Selection Task reasoning for undergraduate students after a 40-hour lecture course in formal logic).

The present study

In the present study we aimed to provide several "best case" tests of the hypothesis that philosophical expertise will diminish the influence of biasing factors on moral judgment and justification – that is, factors that we assume philosophers would not endorse as legitimate upon reflection. We solicited judgments of cases that are widely discussed in the philosophical and psychological literatures, and thus that most philosophers would be familiar with: three variants of the trolley problem, and also the "Asian disease" case introduced by Tversky and Kahneman (1981) to illustrate the effect of "save" versus "die" framing on judgment. (Although the Asian disease case is typically used to illustrate the divergent valuation of gains and losses under prospect theory, of course it depends upon participants making a moral judgment – a decision about the most appropriate course of action when others' lives are at stake.)

We put half of our participants in a "reflection" condition, explicitly instructing them to reflect before making their judgments, imposing a time delay to ensure at least a minimal level of reflection, and specifically encouraging them to consider potential alternative phrasings and variants of each case before submitting their response. At the end of the test, we asked participants whether they endorsed two putative moral principles, including the Doctrine of the

Double Effect. Finally, we asked participants a few questions designed to gauge their level of expertise with the particular scenario types.

Our design allows for several tests of biased response under these conditions. We can assess (1) susceptibility to the effect of "die" versus "save" framings for Asian disease-type cases. We can ask whether order of presentation affects the judgment of (2) trolley-type problems (replicating Schwitzgebel and Cushman 2012) as well as (3) Asian disease-type problems. For trolley-type cases, we can ask whether (4) these order effects carry over to influence the endorsement of putative moral principles such as the Doctrine of the Double Effect.

Also, we varied the specific content of trolley-type cases between participants in order to provide several additional tests. Half of participants viewed the traditional footbridge and sidetrack switch cases, while the other half of participants viewed a modified "drop" case in place of the traditional footbridge case. In the traditional footbridge case, the protagonist pushes his victim off a footbridge with his hands; in the modified "drop" case, the protagonist instead drops his victim via a lever-operated trap door. Thus, the footbridge and switch cases differ both in terms of the Doctrine of the Double Effect and also in the presence of a direct "push" involving physical contact. Both factors have been found to influence moral judgments among ordinary participants (Cushman, Young, and Hauser, 2006). Despite the difference in physical contact, the footbridge and trapdoor cases do not differ in terms of the Doctrine of the Double Effect. This allows us to ask whether (5) order of presentation or exposure to a physical-contact case affects the endorsement of a principle distinguishing cases according to the degree to which harm is caused in a "personal" manner (e.g., a direct push). It also allows us to ask whether (6) the presence of a direct push increases the likelihood of endorsing the Doctrine of the Double

Effect, despite its irrelevance to that doctrine, because it amplifies the divergence in moral judgment between cases that are also distinguished by the Doctrine of Double Effect.

Methods

Participants

We obtained email addresses of potential participants from the websites of philosophy departments and comparison departments in the United States, excluding departments that had been contacted in Schwitzgebel and Cushman (2012). An email invited recipients to participate in a study of philosophers' and similarly educated non-philosophers' judgments about moral dilemmas, linking to a website containing our questionnaire and encouraging recipients to forward the message to academic colleagues.

Near the end of the questionnaire we asked participants' age, nationality, highest degree, highest degree (if any) in philosophy, and "Are you a professor of philosophy?" with response options "yes, and ethics is my area of primary specialization", "yes, and ethics is an area of competence for me", "yes, but not in the area of ethics", or "no". We excluded any participant who did not report having a graduate degree, leaving 497 respondents reporting graduate degrees in philosophy ("philosophers"), 469 (94%) with philosophy PhD's; and 921 respondents reporting graduate degrees, but not in philosophy ("non-philosophers"), 799 (87%) with PhD's. Among the philosophers, 104 (21%) reported being professors of philosophy with specialization in ethics and 167 (34%) reported competence but not specialization. 98% of participants reported U.S. nationality, and 33% reported being female. Age was assessed in ten-year categories, from "Under 15 years", "15 to 24 years", etc., to "65 years and over". The median

response category was "45 to 54 years" for both groups of respondents (only one participant reported age 15-24 and none reported being under 15 years).

Questionnaire design

Reflection vs. control condition. Half of participants were randomly assigned to a reflection condition. Before seeing any scenarios, participants in the reflection condition were told:

> Over the course of the five[1] questions that follow, we are particularly interested in your reflective, considered responses. After each case, please take some time to consider the different moral dimensions at issue, including potential arguments for and against the position to which you are initially attracted. Also please consider how you might respond to different variants of the scenario or to different ways of describing the case. After you finish reading each of the five cases, there will be a 15-second delay to encourage careful reflection before you are asked a question about the case. You needn't answer immediately after the question appears. Please feel free to take as much time as you like.

Also, after each scenario, participants in the reflection condition were told:

> Please take some time to consider the different moral dimensions of the scenario, including potential arguments both for and against [the action described]. Please also consider how you might respond to different variants of the scenario or different ways of describing the case. In fifteen seconds, you will be asked a question about the scenario. You needn't answer immediately after the question

[1] In fact, participants were asked to respond to six questions, not five – an error in the stimulus materials that we did not notice until later.

> appears. We want you to reflect carefully about it, so please take as much time as you like.
>
> When you are ready to BEGIN the reflection period, hit the advance button (>>) below. The text of the scenario will remain on the screen. After 15 seconds you will be permitted to make a response, but take as much time as you would like.

Participants in the control condition were given no special instructions either to answer reflectively or to answer non-reflectively.

Trolley problems. Participants then saw two of three "trolley"-type problems, in random order. One was a *Switch* scenario, involving saving five people in the path of a runaway boxcar by flipping a switch to divert the boxcar onto a sidetrack where it will kill one person. The other was randomly selected to be either a *Push* scenario, involving saving five people by pushing a hiker with a heavy backpack into the path of a runaway boxcar, or a *Drop* scenario, involving saving five people by pulling a lever to drop one person into the path of a runaway boxcar. Respondents rated each scenario on a 1-7 scale from "extremely morally good" (1) through "neither good nor bad" (4) to "extremely morally bad" (7). The exact text of these scenarios and the rest of the questionnaire is available online in the Supplementary Online Materials. We excluded any scenario response that was produced in fewer than 4 seconds (< 1% of responses in the control condition, and by design none in the reflection condition).

We used runaway boxcar scenarios to maximize philosophers' sense of familiarity with the scenario type. Most philosophers, we believe, upon seeing any of the boxcar scenarios, would be swiftly reminded of the famous "trolley problems" of Foot (1967) and Thomson (1985). We hypothesized that philosophers seeing any one of these scenarios might be able to

anticipate the type of scenario that would come next – perhaps especially in the reflection condition, in which we explicitly asked participants to "consider how you might respond to different variants of the scenario". This design thus gave expert participants an excellent chance to reduce the magnitude of any order effect by accurately anticipating the type of scenario that might come next.

We varied Drop and Push, anticipating that participants in Drop might differ less than participants in Push in their endorsements of two abstract principles later in the questionnaire, as we will soon explain.

Framing effect scenarios. Participants then saw two of four loss aversion or framing effect scenarios of the sort made famous by Tversky and Kahneman (1981). In *Save Disease*, an unusual disease is expected to kill 800 people and participants chose between Program A in which 200 people would be saved and Program B in which there was a 1/4 probability that 800 people would be saved and a 3/4 probability that no people would be saved. *Kill Disease* was identical except that the programs were phrased in terms of how many people would die rather than how many would be saved. *Save Nuclear* and *Kill Nuclear* involved a nuclear meltdown expected to kill 600 and probabilities of 1/3 and 2/3. Participants saw either Save Disease and Kill Nuclear or Kill Disease and Save Nuclear, in random order.

The general finding in the literature on loss aversion is that respondents tend to prefer the risky choice (Program B) when the scenario is framed in terms of how many will die and the safe choice (Program A) when the scenario is framed in terms of how many will be saved (Tversky and Kahneman 1981; Kühberger, 1998). We wanted to see if professional philosophers, including professional philosophers explicitly encouraged to consider "different ways of describing the case", and including professional philosophers who regard themselves as experts

on framing effects and loss aversion, would show the same size framing effects as a comparison group. As with the trolley cases, we chose phrasings and cases close to the classic formulations of Tversky and Kahneman so as to give expert participants, especially in the reflection condition, an excellent opportunity to reduce the effects by trying to avoid being excessively swayed by the "saved" vs. "die" phrasing.

Doctrine of the Double Effect and the Personal Principle. Next we asked two questions about moral principles. First, we asked about the famous *Doctrine of the Double Effect:* whether using one person's death as a means of saving others is morally better, worse, or the same as killing one person as a side effect of saving others. Second we asked about a "*Personal Principle":* whether helping several people by harming one person in a personal, face-to-face way is morally better, worse, or the same as helping others by harming one person in a less immediately personal way.

We predicted that philosopher participants who saw Push/Drop before Switch would be more likely to rate the scenarios equivalently and then reject the Doctrine of the Double Effect, as in Schwitzgebel and Cushman (2012). For similar reasons, we predicted that participants would also be more likely to say it's bad to harm in a personal way if they saw Switch before Push than if they saw Push before Switch. Given the generally lower ratings for Push than for Switch, we also predicted that participants in the Push condition would be less likely than those in Drop to say it's better to harm in a personal way. Also, we suspected that participants in the Push condition, if they were less likely to rate the scenario pairs equivalently, might therefore be more likely than those in the Drop condition to endorse a principle that treats the scenarios inequivalently (Doctrine of the Double Effect), despite the apparent irrelevance of the Push-Drop difference to the Doctrine of the Double Effect.

Familiarity, stability, and expertise. Next were the demographic questions already described. Finally, we asked a few questions about familiarity, stability, and expertise. We asked four prior familiarity questions: one concerning trolley problems, one concerning loss aversion/framing effects, one concerning the Doctrine of the Double Effect, and one concerning previous empirical research on philosophers' responses to trolley problems. Respondents who claimed familiarity both with trolley problems and with the Doctrine of the Double Effect were then asked if they regarded themselves as having expertise on those issues and if they regarded themselves as "having had a stable opinion about the trolley problem and Doctrine of Double Effect before participating in this experiment". We asked similar expertise and stability questions for those reporting familiarity with loss aversion/framing effects. Again, see the Supplementary Online Material for exact wording.

Results

Double Effect scenarios

Means. Figure 1 displays the mean results for the Double Effect scenarios. As expected, Push was rated worse than Drop (5.3 vs. 4.5, $t(1398) = 9.3$, $p < .001$, Cohen's $d = 0.48$), which was rated worse than Switch (4.5 vs. 3.7, $t(2094) = 11.1$, $p < .001$, $d = 0.50$). Also as expected, order effects were present for all cases and largest for Switch (Push 5.5 vs. 5.2, $t(697) = 2.4$, $p = .02$, $d = 0.18$; Drop 4.8 vs. 4.3, $t(699) = 4.2$, $p < .001$, $d = 0.31$; Switch 3.2 vs. 4.2, $t(1393) = 12.1$, $p < .001$, $d = 0.62$).

Other predictions were tested with one multiple regression model for each scenario, predicting response from philosopher, reflection condition, presentation in second position, and

all interaction variables. (Here, as in all of the linear and logistic regression models reported in this paper, we code categorical predictor variables as 1 = feature present, -1 = feature absent, and calculate interactions as the product of predictors. This allows us to interpret "main effects" of predictors and their interactions equivalently to an analysis of variance).

If philosophers are less subject to order effects than are non-philosophers, we would expect to see an interaction effect of philosopher by position. If philosophers are less subject to order effects specifically in the reflection condition, we would expect to see a three-way interaction between philosopher, condition, and position. Neither interaction was found for any of the three scenarios, despite sufficient statistical power to detect effects of a "small" size $f^2 >$ .02 (corresponding to partial $r > .15$, Cohen 1988) with at least 95% probability in each case. For Push, the statistically significant predictors were presentation in the second position ($\beta = -0.09$, $t(698) = -2.4$, $p = .02$, $f^2 = .008$), reflection condition ($\beta = 0.09$, $t(698) = 2.2$, $p = .03$, $f^2 = .007$), and philosopher respondent ($\beta = -0.08$, $t(698) = -2.0$, $p = .046$, $f^2 = .006$). For Drop, the only significant predictors were reflection condition ($\beta = 0.21$, $t(700) = 5.5$, $p < .001$, $f^2 = .044$) and position ($\beta = -0.15$, $t(700) = -3.9$, $p < .001$, $f^2 = .023$). For Switch, the significant predictors were position ($\beta = 0.30$, $t(1394) = 11.2$, $p < .001$, $f^2 = .091$) and philosopher respondent ($\beta = -0.06$, $t(1394) = -2.4$, $p = .02$, $f^2 = .004$).

As a manipulation check, we confirmed that response time in the reflection condition exceeded that in the control condition. In the control condition the median response time was 49 seconds for the first scenario and 34 seconds for the second scenario. In the reflection condition, median response times were about double: 98 and 66 seconds respectively.

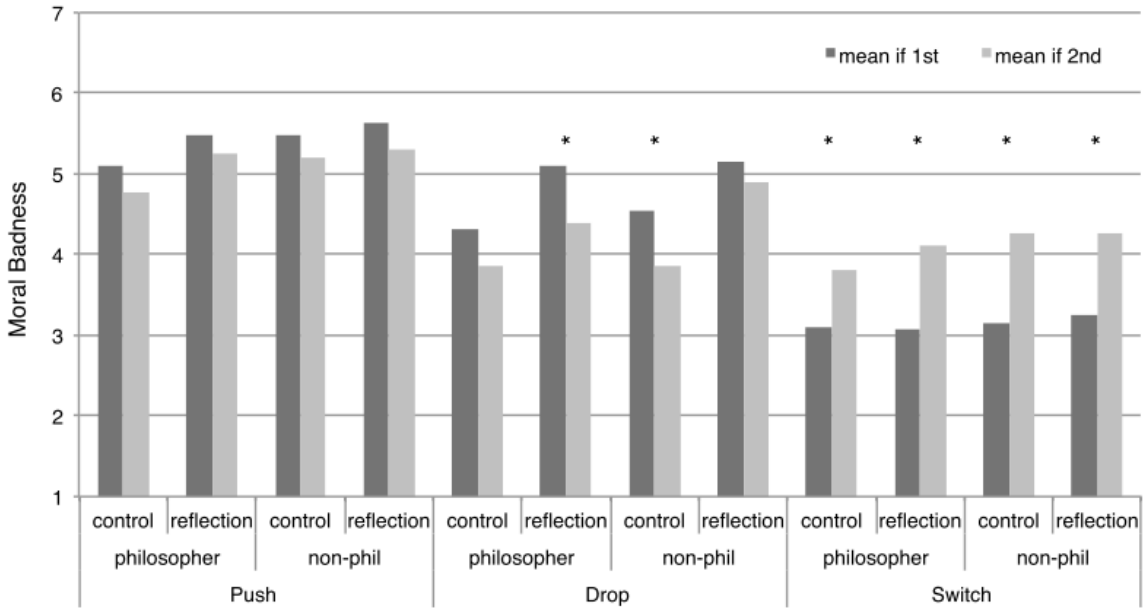

Figure 1: Trolley problem mean moral ratings by order of presentation, non-reflection (control) vs. reflection conditions, and professional philosophers vs. non-philosophers. Higher values indicate greater moral badness. Asterisks indicate one-tailed statistical significance at $p < .05$ for each pair of adjacent bars. For aggregated statistics see the main text.

Equivalency ratings. We also analyzed equivalency ratings. Participants were coded as having rated the scenario pairs *equivalently* if they gave both scenarios the same 1-7 rating, *inequivalently* if they rated the scenarios differently in the predicted direction (that is, Push or Drop worse than Switch), and they were excluded if they rated the scenarios differently in the unpredicted direction (that is, Switch worse than Push or Drop: 2% of participants). Equivalency ratings are less subject to scaling concerns and correspond more closely to canonical statements of the Doctrine of the Double Effect, which is generally expressed in terms of the inequivalency of harm as a means vs. harm as a side-effect.

Figure 2 shows the equivalency results. As predicted from results in Schwitzgebel and Cushman (2012), respondents were more likely to rate the scenarios equivalently if Push or Drop was presented before Switch, since ratings in the Switch scenario tend to be labile and matched to the first-presented scenario if Switch is presented second. Push and Switch were rated equivalently by 24% of respondents when Switch was presented first and 45% of respondents when Switch was presented second (Fisher's exact, $N = 689$, $p < .001$, odds ratio (OR) = 0.39). Drop and Switch were rated equivalently by 46% of respondents when Switch was presented first and 70% of respondents when Switch was presented second (Fisher's exact, $N = 671$, $p < .001$, OR = 0.37).

Other predictions were tested by a binary logistic regression model, predicting response from Drop condition, philosopher, reflection condition, Switch-first condition, and all interaction variables. If philosophers are less subject to order effects than are non-philosophers, we would expect to see an interaction effect of philosopher by Switch-first. If philosophers are less subject to order effects specifically in the reflection condition, we would expect to see a three-way interaction between philosopher, reflection condition, and Switch-first. For both interaction

effects our analysis had a power of 95% to detect an odds ratio of 1.23 (or its reciprocal, 0.81). No interaction effects were statistically significant. Significant predictors were Switch first (OR = 0.63, $p < .001$), Drop condition (OR = 1.7, $p < .001$), and reflection condition (OR = 0.72, $p < .001$). Thus, although participants in the reflection condition were less likely in general to rate the scenarios equivalently (39% vs. 54%, Fisher's exact, N = 1360, $p < .001$), we found no evidence of the interaction between reflection condition and order that would be expected if being asked to reflect reduced the order effects. As predicted, we observed higher equivalency in the Drop condition and lower equivalency in the Switch-first condition. A model with those three predictive variables (Switch first, Drop condition, reflection condition) plus philosopher and philosopher-by-Switch-first yields a non-significant trend toward smaller order effects for philosophers (OR = 1.1, $p = .08$, CI 0.99 to 1.3).

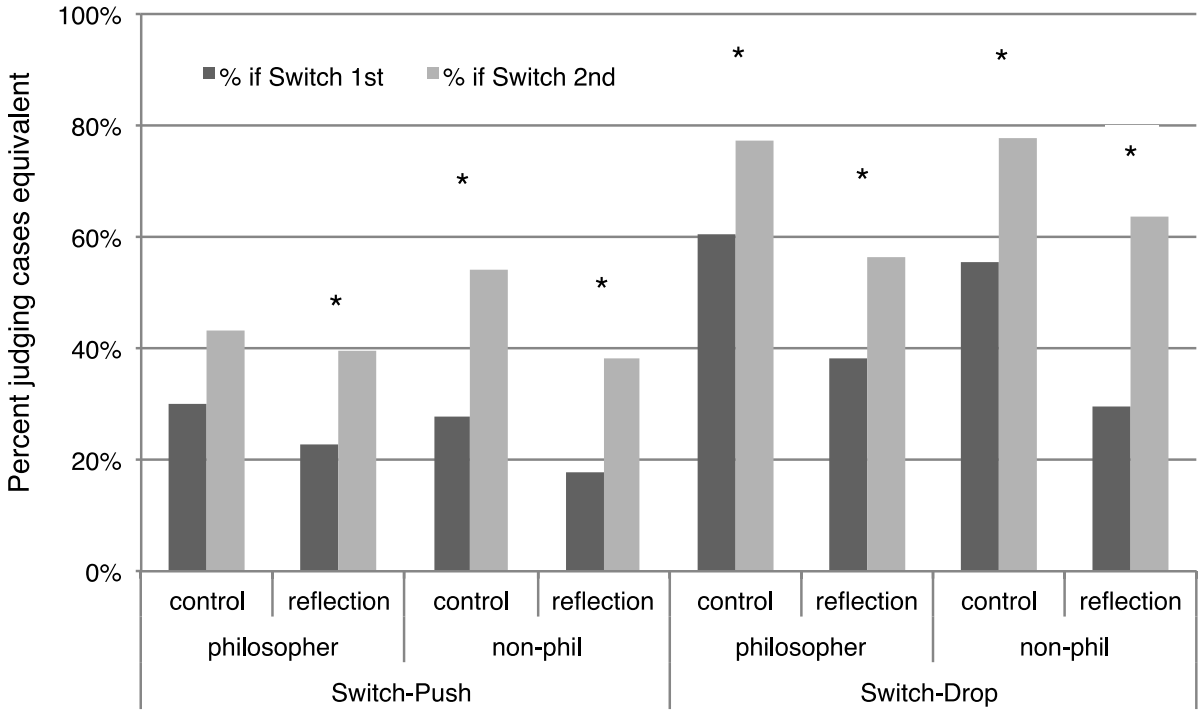

Figure 2: Percentage of participants rating the two trolley problems equivalently, by order of presentation, non-reflection (control) vs. reflection conditions, and professional philosophers vs. non-philosophers. Asterisks indicate one-tailed statistical significance at $p < .05$ for each pair of adjacent bars.

Framing effect scenarios

The disease and nuclear scenarios differed only slightly in overall percentage favoring the risky choice (57% vs. 62%, N = 2732, p = .007, OR = 0.81) and did not detectably differ in the size of the framing or order effects, so the two scenario types were merged for analysis. Figure 3 displays the results.

Median response time in the control condition was 45 seconds for the first-presented scenario and 33 seconds for the second-presented scenario. In the reflection condition, median response times were a bit less than double: 77 seconds and 59 seconds, respectively.

Framing effects. In the first-presented scenario, as Tversky and Kahneman (1981) and most subsequent studies have found, participants were much more likely to select the risky option (Program B) when the options were expressed in terms of how many of the disaster victims expected to die will "die" than when an otherwise equivalent pair of options was presented in terms of how many will "be saved". (To see the traditional framing effect in the figure, look only at the dark bars in the graphs which represent the first-presented scenarios compared between participants.) The effect was large both for philosophers (79% vs. 32%, Fisher's exact, N = 475, p < .001, OR = 7.9) and for non-philosophers (83% vs. 43%, Fisher's exact, N = 903, p < .001, OR = 6.4).

Other predictions were tested by a binary logistic regression model, predicting first-scenario response from "die" frame, philosopher, reflection condition, and all interaction variables. Significant predictors were frame (OR = 2.7, p < .001) and philosopher respondent (OR = 0.84, p = .008). In a model with frame, philosopher, and frame-by-philosopher, philosophers showed nominally larger framing effects, but this effect did not approach

significance: interaction OR = 1.1, p = .41, CI 0.93 to 1.2). This analysis had a power of 95% to detect an odds ratio of 0.81 (or its reciprocal, 1.23).

Order effects. To see the order effects in Figure 3, compare the pairs of adjacent dark and light bars. In every case, preference for the risky option is significantly closer to 50% when the scenario is presented second, having followed presentation of a very similar scenario with the opposite framing, than when the scenario was presented first.

Other predictions were tested by two binary logistic regression models, predicting "die"-frame response and "save"-frame response from philosopher, reflection condition, second-position presentation, and all interaction variables. If philosophers are less subject to order effects than are non-philosophers, we would expect to see an interaction effect of philosopher by position. If philosophers are less subject to order effects specifically in the reflection condition, we would expect to see a three-way interaction between philosopher, reflection condition, and position. No interaction variable was significant in either model. In both analyses, responses were closer to 50% in the second position ("die" frame: OR = 0.59, $p < .001$; "save" frame: OR = 1.5, $p < .001$) and philosophers were less likely to favor the risky choice ("die" frame OR = 0.84, p = .007; "save" frame OR = 0.85, p = .005). Models with position, philosopher, and position-by-philosopher show nominally larger interaction order effects for philosophers that do not approach statistical significance: ("die": OR = 0.95, p = .38, CI 0.83 to 1.07; "saved": OR = 1.1, p = .22, CI 0.96 to 1.20). For all the interactions reported above, our analysis had a power of at least 95% to detect an odds ratio of 0.80 (or its reciprocal 1.25).

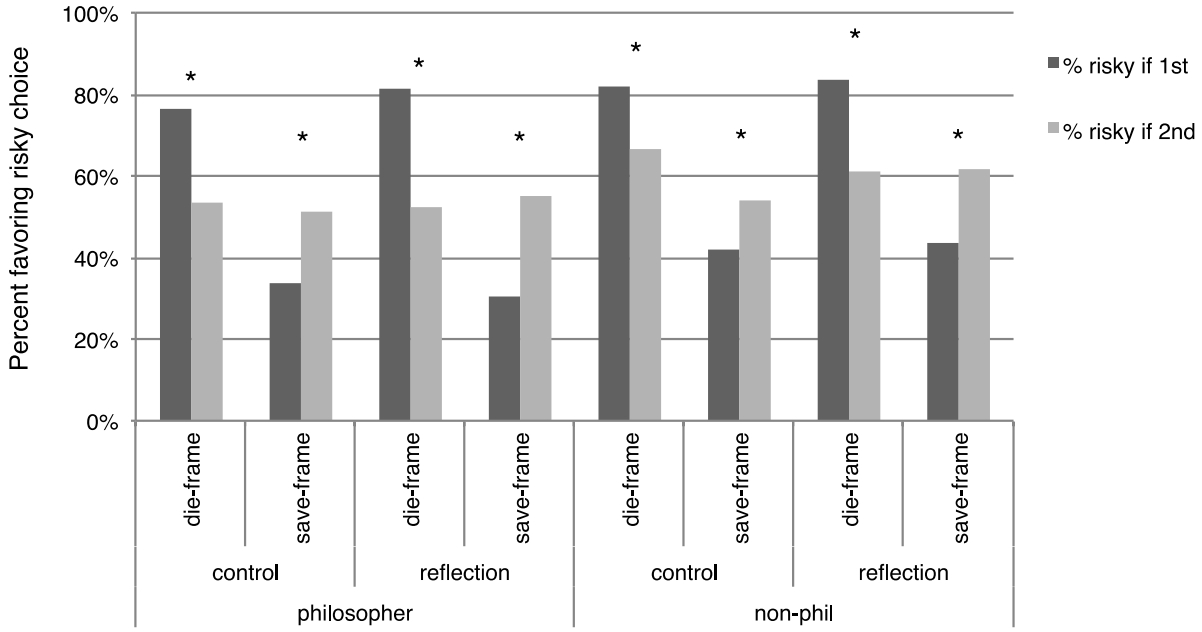

Figure 3: Percentage of participants favoring the risky choice in "die" vs. "saved" framing scenarios, by order of presentation, non-reflection (control) vs. reflection conditions, and professional philosophers vs. non-philosophers. Asterisks indicate one-tailed statistical significance at $p < .05$ for each pair of adjacent bars.

Endorsement of Principles

We did not see the expected order effects on endorsement for either the Doctrine of the Double Effect or the Personal Principle. Philosophers' DDE endorsement (worse to harm as means than as side effect) was 59% with Switch first vs. 62% with Push or Drop first (Fisher's exact, N = 462, p = .57, CI for diff -12% to +6%). Non-philosophers were actually *more* likely to endorse DDE if they received Push or Drop before Switch: 51% vs. 58% (Fisher's exact, N = 845, p = .045, OR = 0.75) – a result for which we have no explanation, but which we also found in Schwitzgebel and Cushman (2012). Only 66 philosopher participants unfamiliar with previous research on philosophers' responses to trolley problems were in the Push version of the control condition – the condition closest to the original Schwitzgebel and Cushman (2012). Unfortunately, this is too few to allow an appropriately powered direct comparison with that study's finding of a 62% vs. 46% order effect on endorsement of DDE. (The confidence interval for the effect size in this subgroup does include the observed 16% effect size in our earlier work: 22/40 vs. 14/26, CI for difference -26% to +23%, Z = 0.1, p = .93.)

For the Personal Principle philosophers split 26% better, 63% same, 11% worse when Switch was first vs. 27%-59%-14% when Push or Drop was first (χ^2= 1.4, p = .49; φ = .05); for non-philosophers it was 26%-61%-13% vs. 30%-61%-10% (χ^2 = 3.3, p = .19; φ = .06). We did find that philosophers' endorsements of the Personal Principle were substantially influenced in the predicted direction by whether they had been assigned to the Push or Drop condition. In the Drop condition, 32% of philosophers said harm done in a personal manner was morally better than harm done impersonally, compared to 22% in the Push condition (58% vs. 63% said "same", and 10% vs. 16% said worse, χ^2 = 8.4, p = .02, φ = .13). In contrast, non-philosophers showed no detectable effect (30%-59%-10% vs. 25%-62%-12%, χ^2 = 3.2, p = .20, φ = .06),

consistent with Schwitzgebel and Cushman's (2012) finding that philosophers were more likely than non-philosophers to shift their endorsements of principles to match their experimentally manipulated judgments about scenarios.

Finally, we found some evidence that endorsements of the Doctrine of the Double Effect were also influenced by whether participants were assigned to the Push or Drop condition. As we saw above, participants who viewed Drop were more likely to rate it equivalent to Switch than were participants who viewed Push. Both pairs of cases differ along a dimension captured by the DDE; participants' higher likelihood of rating the Push-Switch pair inequivalently than the Drop-Switch pair is likely a consequence of the additional presence of an up-close, personal harm in Push. We reasoned that participants might exhibit greater endorsement of the DDE as a convenient explanation for their discrepant judgments of Push and Switch than for their more weakly discrepant judgments of Drop and Switch cases. This effect is of particular interest because it involves the misapplication of a judgment driven by one stimulus feature (an up-close, personal harm) to the endorsement of another stimulus feature (harm caused as a means to an end). Consistent with this predicted effect we found that non-philosophers were more likely to endorse the DDE when they had seen Push (60%) than if they had seen Drop (50%; Fisher's exact, $N = 845$, $p = .005$, $OR = 1.5$). We found a non-significant trend in the same direction for philosophers (63% vs. 57%; Fisher's exact, $N = 462$, $p = .18$, $OR = 1.3$).

Familiarity, Stability, and Expertise

A majority of philosophers reported familiarity with the types of cases we used, and a sizable minority claimed expertise and stability of opinion (Table 1). Philosophers reporting familiarity, expertise, stability, and specialization in ethics appeared to be just as subject to order

effects as did philosophers reporting unfamiliarity, lack of expertise, lack of stability, and lack of specialization in ethics (Figures 4 and 5). As is evident from the figures, philosophers reporting familiarity, expertise, stability, and specialization in ethics trended toward showing *larger* order effects than the remaining philosophers. For example, among philosopher respondents reporting being philosophy professors with an area of specialization in ethics, 26% rated the scenarios equivalently when Switch was first vs. 56% when Push or Drop was first (Fisher's exact, N = 99, p = .004, OR = 0.27), compared to a 42%-52% shift for all remaining philosopher respondents (Fisher's exact, N = 376, p = .06, OR = 0.67). However, these trends are non-significant in binary logistic regressions (e.g., interaction effect of order and specialization: OR = 1.3, p = .06). Due especially to the multiple comparisons included in this set of analyses, we must interpret these results with caution.

Due to the smaller sample size, we had less statistical power to detect significant differences in the eleven comparisons charted in Figures 4 and 5, which contrast subgroups of philosophers, than for our comparisons between philosophers and non-philosophers. For these comparisons, we had a power, at $\beta = .80$, to detect odds ratios from 1.3 to 1.5 (or their reciprocals, 0.76 to 0.68), depending on the specific comparison.

Table 1: Percentage of respondents claiming familiarity, expertise, or stability of opinion on the trolley problem and the Doctrine of the Double Effect, framing effects and loss aversion, and empirical studies of philosophers' responses to trolley problems.

		trolley problems and Double Effect	framing effects and loss aversion	empirical studies of philosophers
phil	familiarity	77%	62%	33%
	expertise	20%	13%	--
	stability	40%	26%	--
non-phil	familiarity	9%	24%	4%
	expertise	1%	4%	--
	stability	3%	9%	--

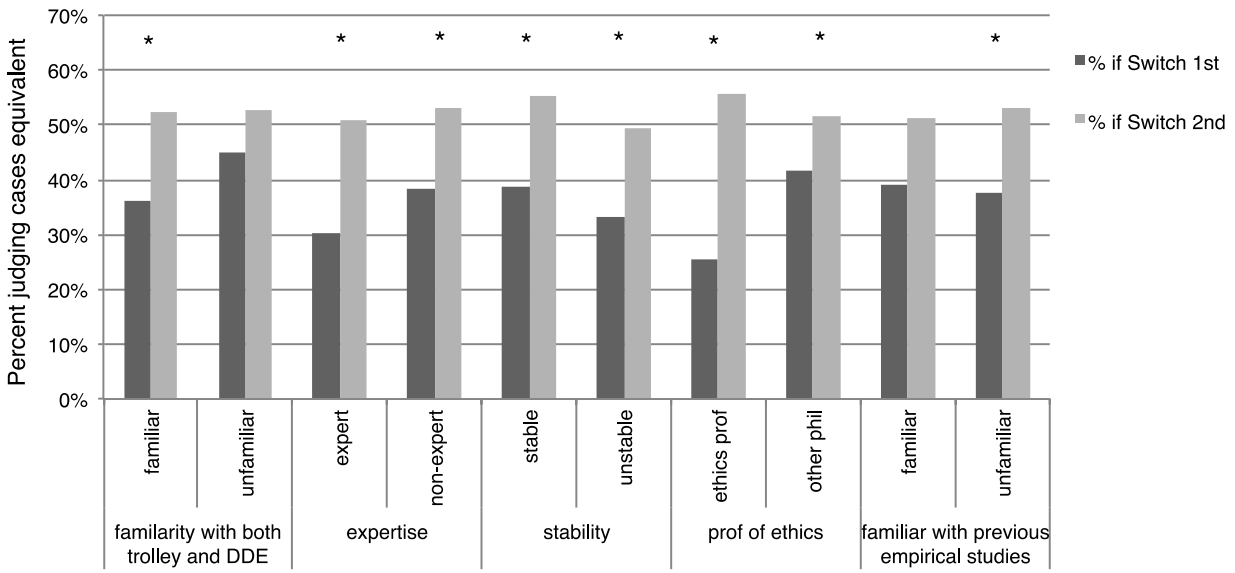

Figure 4: Percentage of philosophers rating the two trolley problems equivalently, by order of presentation, broken down by types of expertise. Asterisks indicate one-tailed statistical significance at $p < .05$ for each pair of adjacent bars.

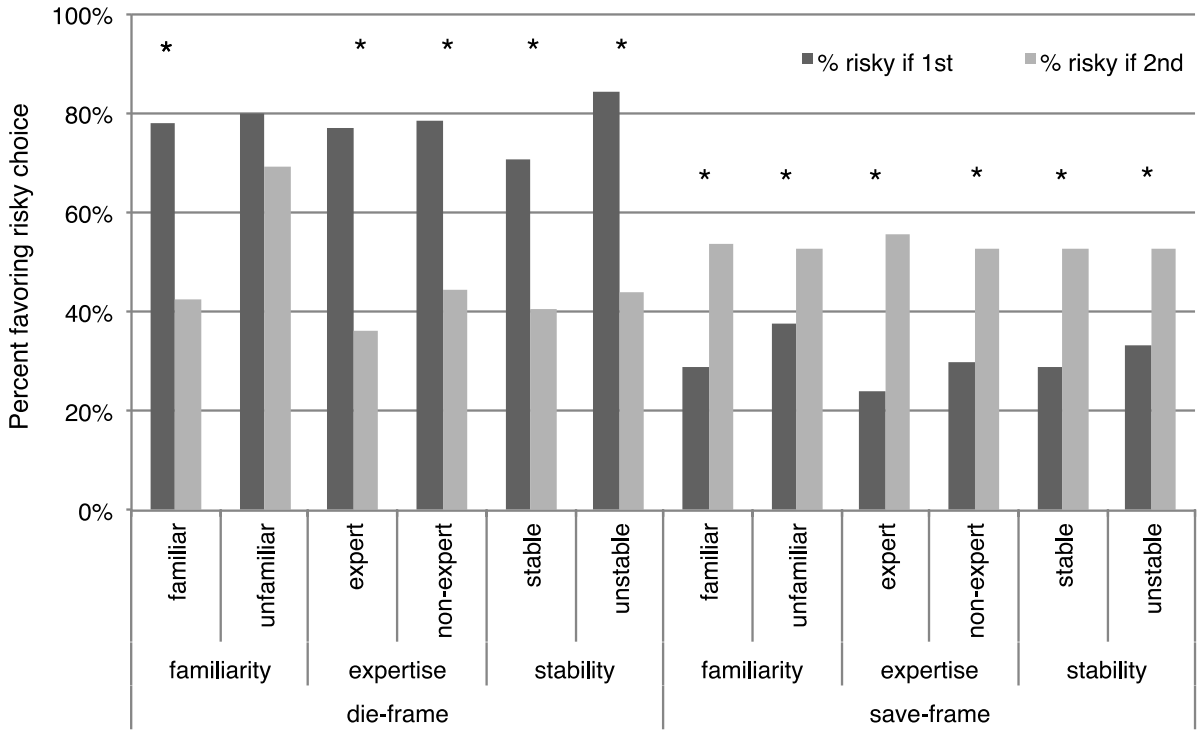

Figure 5: Percentage of philosophy participants favoring the risky choice in "die" vs. "saved" framing scenarios, by order of presentation and level of familiarity or expertise. Asterisks indicate one-tailed statistical significance at $p < .05$ for each pair of adjacent bars.

Discussion

Replicating prior research, we found substantial order effects on participants' judgments about the Switch version of trolley problem, substantial order effects on their judgments about making risky choices in loss-aversion-type scenarios, and substantial framing effects on their judgments about making risky choices in loss-aversion-type scenarios.

Moreover, we could find no level of philosophical expertise that reduced the size of the order effects or the framing effects on judgments of specific cases. Across the board, professional philosophers (94% with PhD's) showed about the same size order and framing effects as similarly educated non-philosophers. Nor were order effects and framing effects reduced by assignment to a condition enforcing a delay before responding and encouraging participants to reflect on "different variants of the scenario or different ways of describing the case". Nor were order effects any smaller for the majority of philosopher participants reporting antecedent familiarity with the issues. Nor were order effects any smaller for the minority of philosopher participants reporting expertise on the very issues under investigation. Nor were order effects any smaller for the minority of philosopher participants reporting that before participating in our experiment they had stable views about the issues under investigation.

Previous research has found substantial loss-aversion framing effects even among fairly sophisticated participants (reviewed in Kühlberger 1998; Reyna et al. 2014). The present study confirms and extends these results to very high levels of expertise. That the effect is present in participants with very high levels of expertise raises the question of whether those experts might in fact be responding rationally to relevant factors, contrary to our initial assumptions in experimental design. For example, Mandel (2014) argues that participants might naturally read

"200 people will be saved" as meaning something like *at least* 200 people will be saved (and maybe more), and comparably "400 people will die" as meaning something like *at least* 400 people will die – in which case it might be rational to prefer the risky choice in the die frame and the safe choice in the save frame. If Mandel's explanation were correct in the present case, however, we might expect to see the same frame-driven pattern in the second-presented scenarios as in the first-presented scenarios, since the wording is the same; and we would probably expect to see smaller framing effects among expert participants who were presumably aware that the intended interpretation of the options is exact numbers saved and dying, not minimum numbers. It remains open, however, that there are other ways of interpreting the framing effects and order effects as rational, contra existing psychological orthodoxy.

Our results cast doubt on some commonsense approaches to bias reduction in scenario evaluation: training in logical reasoning, encouraging deliberative thought, exposure to information both about the specific biases in question and about the specific scenarios in which those biases manifest. Future efforts to minimize cognitive bias might more effectively focus on other means, such as alterations of choice architecture (Thaler & Sunstein 2007) or feedback-based training and social support (Mellers et al. 2014).

Our findings on the effect of contextual factors on philosophers' endorsement of moral principles were more equivocal. We found that assignment to different pairs of trolley cases (Drop-Switch, not differing in degree of personal contact between agent and victim, vs. Push-Switch, very different in degree of personal contact between agent and victim) substantially influenced philosophers' endorsements of a principle regarding the value or disvalue of harming in a personal face-to-face way. We found a significant effect of case on endorsement of the Doctrine of the Double Effect for non-philosophers. However, we did not find a significant

case-on-endorsement effect for philosophers, nor did we replicate Schwitzgebel and Cushman's (2012) finding that the order of presentation of trolley-type dilemmas affects philosophers' subsequent endorsement of the Doctrine of the Double Effect. Such mixed results are consistent with Schwitzgebel and Cushman's finding that philosophers' endorsements of abstract principles are substantially influenced by contextual factors, such as perhaps in this case the presence of the loss-aversion cases, the absence of the moral luck and action-omission cases, and the assignment of half of the participants into a reflection condition. Further research would help to clarify the effect of order, case, and similar spurious factors on philosophers' endorsement of moral principles. Further research would also clarify the extent to which philosophers' assessments of non-moral cases and principles – for example, in philosophy of mind or in formal logic – are subject to the same types of effects.

We confess that we find our main result surprising: that is, our across-the-board failure to find evidence for philosophical expertise and reflection in moderating biased moral judgment. We would have thought – we are still inclined to think – that at a high-enough level of expertise people won't be much swayed by order and framing, at least when they report having stable opinions and are encouraged to answer reflectively. We wouldn't expect, for example, that Judith Jarvis Thomson (1985) or John Martin Fischer (Fischer & Ravizza, 1992) would rate Push and Switch equivalently if the scenarios are presented in one order and inequivalently if they are presented in another order. However, if there is a level of philosophical expertise that reduces the influence of factors such as order and frame upon one's moral judgments, we have yet to find empirical evidence of it.

Acknowledgements

We thank Daniel Guevara, David Mandel, Regina Rini, and Walter Sinnott-Armstrong for their valuable suggestions regarding this manuscript, and members of the Moral Psychology Research Laboratory for their assistance with data collection. We gratefully acknowledge the University of California, Riverside Academic Senate for its financial support.

References

Baron, J. 2000: *Thinking and deciding*. Cambridge University Press.

Bennett, J. 1998: *The Act Itself.* Oxford: Clarendon.

Buckwalter, W. forthcoming. Intuition fail: Philosophical activity and the limits of expertise. *Philosophy & Phenomenological Research.*

Cheng, P.W., Holyoak, K.J., Nisbett, R.E., & Oliver, L.M. 1986: Pragmatic versus syntactic approaches to training deductive reasoning. *Cognitive Psychology, 18,* 293-328.

Cohen, J. 1988: *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cushman, F., Young, L., & Hauser, M. 2006: The role of conscious reasoning and intuition in moral judgment. *Psychological Science, 17,* 1082-1089.

Fischer, J. M., & Ravizza, M. 1992: *Ethics: Problems and principles*. New York: Holt, Rinehart & Winston.

Foot, P. 1967: The problem of abortion and the doctrine of double effect. *Oxford Review,* 5, 5-15.

Greene, J. (2014). *Moral tribes: emotion, reason and the gap between us and them*. Atlantic Books.

Grundmann, T. 2010: Some hope for intuitions: A reply to Weinberg. *Philosophical Psychology,* 23, 481-509.

Heijltjes, A., van Gog, T., Leppink, J., & Paas, F. 2014: Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction, 29,* 31-42.

Howard-Snyder, F. 2002/2011: Doing vs. allowing harm. *Stanford Encyclopedia of Philosophy* (Winter 2011 edition).

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kühberger, A. 1998: The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75,* 23-55.

Kuhn, D. 1991: *The skills of argument.* Cambridge: Cambridge University Press.

Lehman, D.R., Lempert, R.O., & Nisbett, R.E. 1988: The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43,* 431-442.

Liao, S.M., Wiegmann, A., Alexander, J., & Vong, G. 2012: Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology, 25,* 661-671.

Livengood, J., Sytsma, J., Feltz, A., Scheines, R., & Machery, E. 2010: Philosophical temperament. *Philosophical Psychology, 23,* 313-330.

Ludwig, K. 2007: The epistemology of thought experiments: First person versus third person approaches. *Midwest Studies in Philosophy, 31,* 128-159.

Machery E. 2011: Expertise and intuitions about reference. *Theoria, 73,* 37-54.

Mandel, D.R. 2014: Do framing effects reveal irrational choice? *Journal of Experimental Psychology: General, 143,* 1185-1198.

McIntyre, A. 2004/2011: Doctrine of double effect, *Stanford Encyclopedia of Philosophy* (Fall 2011 edition).

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., ... & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological science*, *25*, 1106-1115.

Mizrahi, M. 2015: Three arguments against the expertise defense. *Metaphilosophy, 46,* 52-64.

Nado, J. forthcoming: Philosophical expertise and scientific expertise. *Philosophical Psychology.*

Nagel, T. 1979: *Mortal Questions*. Cambridge: Cambridge University Press.

Nelkin, D. K. 2004/2013: Moral luck. *Stanford Encyclopedia of Philosophy* (Winter 2013 edition).

Quinn, W. S. 1989: Actions, intentions, and consequences: the doctrine of doing and allowing. *The Philosophical Review*, 145, 287-312.

Reyna, V.F., Chick, C.F., Corbin, J.C., & Hsia, H.N. 2014: Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students. *Psychological Science 25,* 76-84.

Rini, R.A. 2015: How not to test for philosophical expertise. *Synthese, 192,* 431-452.

Ritchhart, R., & Perkins, D.N. 2005: Learning to think: The challenges of teaching thinking. In K.J. Holyoak & R.J. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning.* Cambridge: Cambridge.

Schulz, E., Cokely, E.T., & Feltz, A. 2011: Persistent bias in expert judgments about free will and moral responsibility: A test of the expertise defense. *Consciousness & Cognition 20,* 1722-1731.

Schwitzgebel, E., & Cushman, F. 2012: Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27,* 135-153.

Sinnott-Armstrong, W. 2008: Framing moral intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, vol. 2: The Cognitive Science of Morality.* Cambridge, MA: MIT.

Thomson, J. J. 1985: The trolley problem. *The Yale Law Journal, 94,* 1395-1415.

Tobia, K., Buckwalter, W., & Stich, S. 2013: Moral intuitions: Are philosophers experts? *Philosophical Psychology, 26,* 629-638.

Tobia, K., Chapman, G., & Stich, S. 2013: Cleanliness is next to morality, even for philosophers. *Journal of Consciousness Studies, 20* (no. 11-12), 195-204.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Tversky, A., & Kahneman, D. 1981: The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

Tversky, A., & Kahneman, D. 1983: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90,* 293-315.

Weinberg, J., Gonnerman, C., Buckner, C., & Alexander, J. 2010: Are philosophers expert intuiters? *Philosophical Psychology,* 23, 331-355.

Williams, B. 1981: *Moral Luck*. Cambridge: Cambridge University Press.

Williamson, T. 2011: Philosophical expertise and the burden of proof. *Metaphilosophy, 42,* 215-229.

Wright, J. 2010: On intuitional stability: The clear, the strong, and the paradigmatic. *Cognition, 115,* 491-503.

Wright, J. 2013: Tracking instability in our philosophical judgments: Is it intuitive? *Philosophical Psychology, 26,* 485-501.