RUNNING HEAD: Expertise in Moral Reasoning?

# Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers

Eric Schwitzgebel
Department of Philosophy
University of California at Riverside

Fiery Cushman
Department of Psychology
Harvard University

Word count (excluding title, abstract, tables, and bibliography): 6,115 words

# Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers

Abstract

We examined the effects of order of presentation on the moral judgments of professional philosophers and two comparison groups. All groups showed similar-sized order effects on their judgments about hypothetical moral scenarios targeting the doctrine of the double effect, the action-omission distinction, and the principle of moral luck. Philosophers' endorsements of related general moral principles were also substantially influenced by the order in which the hypothetical scenarios had previously been presented. Thus, philosophical expertise does not appear to enhance the stability of moral judgments against this presumably unwanted source of bias, even given familiar types of cases and principles.

Keywords: morality, reasoning, cognitive ability, social cognition, experimental philosophy, intuition

**Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers**

## 1. Introduction

Moral judgment is sometimes claimed to arise mostly from automatic processes that depend little on conscious reasoning from general principles (Haidt, 2001; Mikhail 2009). Recent work in moral psychology suggests that people can have trouble explaining the bases of their moral judgments (Cushman, Young, & Hauser, 2006; Haidt & Hersh, 2001; Wheatley & Haidt, 2005) and that moral judgments are influenced by factors most people would deem irrelevant, such as the presence of an odor (Schnall, Haidt, Clore, & Jordan, 2008), the presence or absence of direct physical contact (Cushman, et al., 2006), or the order in which hypothetical moral scenarios are

We thank John Fischer, Joshua Knobe, Joe Paxton, Michael Ridge, Gary Watson, and Liane Young for their valuable contributions in the preparation of this manuscript. We gratefully acknowledge the University of California, Riverside Academic Senate and the Harvard Mind/Brain/Behavior Initiative for their support.

Address for correspondence: Eric Schwitzgebel, Department of Philosophy, University of California at Riverside, Riverside, CA 92521, USA

Email: eschwitz at domain: ucr.edu

The authors contributed equally to this research.

presented (Lombrozo, 2009; Petrinovich & O'Neill, 1996). On the basis of such considerations it is sometimes claimed that, when participants are asked to describe the general principles underlying their moral judgments, they are engaged mostly in post-hoc rationalization disconnected from the real psychological bases of their judgments (Ditto & Liu, in press; Haidt, 2001). Psychologists have by no means reached consensus on this point, however, with others arguing that moral reasoning does often influence moral judgment (Bartels, 2008; Cushman, et al., 2006; Paxton & Greene, 2010; Pizarro & Bloom, 2003; a view also reflected in Kohlberg, 1984).

Because of their extensive training, professional philosophers are a 'best case' population for the skillful use of principled reasoning to influence moral judgment, and they have occasionally been explicitly described as such by psychologists (e.g., Haidt 2001, p. 819 and 829). Indeed, professional ethicists sometimes describe themselves as experts at moral reasoning (Crosthwaite, 1995; Føllesdal, 2004; Singer, 1972). And in reaction to critiques by Jonathan Weinberg and others (e.g., Weinberg, Gonnerman, Buckner, & Alexander, 2010), a number of philosophers have recently asserted that their professional training helps protect them from unconscious and unwanted biases in their domain of expertise (Grundmann, 2010; Hofmann, 2010; Horvath, 2010; Williamson, in press; Wright, 2010).

There is some empirical cause for optimism about philosophical expertise in moral reasoning. Rest (1993) found that people with graduate training in philosophy responded with more sophistication than did non-philosophers to moral dilemmas, like Kohlberg's (1984) famous 'Heinz' dilemma about stealing a drug to save a life. Kuhn (1991) found that philosophy graduate students treated evidence and argument, in general, more skillfully than did other groups. Livengood and colleagues (2010) found that philosophers were more likely than non-

philosophers to succeed at the Cognitive Reflection Test, a series of conceptually tricky but computationally simple math problems (Frederick, 2005).

It remains possible, however, that philosophers' apparent skill at moral argumentation is mostly skill at the post-hoc rationalization of judgments driven by automatic processes that would not necessarily be endorsed upon explicit reflection. Recent evidence suggesting that professional ethicists behave similarly to others of similar social background (e.g., in voting rates, in courtesy at conferences, in rates of charitable giving, in rates of National Socialist party membership during the Nazi era, and in peer-evaluated overall moral behavior) may support the post-hoc rationalization view, if we assume that non-rationalizing philosophical moral reflection would tend to precipitate changes in behavior (Leaman, 1993; Schwitzgebel & Rust, 2009, 2010, in preparation; Schwitzgebel, Rust, Moore, Huang, & Coates, in press).

We test whether philosophical expertise enhances stable reasoning from moral principles by examining order effects on moral judgments about hypothetical scenarios and also on the endorsement of general moral principles. We compare these effects among professional philosophers, non-philosopher academics, and non-academics. We assume that few people—philosophers or non-philosophers—think that one ought to judge case A worse than case B when judged in the order A-B but not when judged in the order B-A. Similarly, we assume that few people think that such variations of presentation order ought to subsequently affect the endorsement of a general principle governing cases A and B. Thus, to the extent moral judgment derives from stable general principles, it should be insulated from such order effects.

Ordinary non-philosophers do show order effects upon their moral judgments (Lombrozo, 2009; Petrinovich & O'Neill, 1996). Our question is whether professional philosophers are any less subject to such order effects. If so, it might warrant optimism about

philosophical expertise and support a model of moral cognition on which skill at explicit moral reasoning can help protect people from unwanted influences on their moral judgments. If not, it would suggest a more pessimistic view about the power of explicit moral reasoning to protect against unwanted sources of bias, even in a best-case population. It would also raise a practical concern about the security of the intuitions that ground philosophical inquiry.

We targeted three moral principles drawn from the philosophical literature, chosen because they are well-known among philosophers and also exhibited in non-philosophers' moral judgments. According to the *doctrine of the double effect,* it is worse to harm a person as a means of saving others than to harm a person as a side-effect of saving others (Foot, 1967; McIntyre, 2004/2009; Thomson, 1985). This principle is illustrated by the famous 'trolley problem': Many people consider it morally worse to throw somebody in front of a train as a means of stopping it from hitting five others (the 'push' case) than to divert a train away from five people and, as a side-effect, hit one person instead (the 'switch' case) (Hauser, et al., 2007; Mikhail, 2000; Petrinovich & O'Neill, 1996).

According to the principle of *moral luck,* we can be morally assessable for outcomes partly outside our control (Nagel, 1979; Nelkin, 2004/2008; Williams, 1981). For example, a reckless driver who kills a pedestrian may deserve more punishment than one who does not, even if the difference in outcome was a matter of chance. Non-philosophers' moral judgments often accord with moral luck (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009; Gino, Shu, & Bazerman, 2010; Young, Nichols, & Saxe, 2010).

A third principle targeted the difference between killing versus letting die—for example, causing someone to drown versus not saving her from drowning. Such cases have been invoked to support a moral principle distinguishing between *action and omission* or doing and allowing

(Bennett, 1998; Howard-Snyder, 2002/2007; Quinn, 1989), again reflected in ordinary people's judgments (Baron & Ritov, 2004; Cushman, et al., 2006; Spranca, et al., 1991).

We presented participants with hypothetical scenarios varying along the dimensions indicated by these moral principles, varying the order of presentation between subjects. If participants are stably applying (or declining to apply) the three moral principles, their response patterns should reflect (or fail to reflect) those principles independently of order of presentation. Finally, we asked participants whether they endorsed each of these principles in general form. If participants stably embrace general principles rather than merely recruiting general principles post-hoc to rationalize prior judgments, the order of presentation of the scenarios should have little influence on subsequent endorsement of the general principles.

## 2. Methods

### 2.1 Participants

We surveyed participants using the Moral Sense Test website (http://moral.wjh.harvard.edu), recruiting through direct emails to philosophy and non-philosophy departments at 25 major research universities with well-ranked PhD programs in philosophy (a minority of participants were recruited through academic blogs). Our usable sample comprised 324 'philosophers' (completed MA or PhD in philosophy), 753 'academic non-philosophers' (completed Master's or PhD not in philosophy), and 1389 'non-academics' (no Master's or PhD in any field) tested between October 2008 and July 2009. Among philosophers, 221 (68%) claimed an area of specialization or competence in ethics and 91 of those also claimed a PhD ('ethics PhDs').

We excluded 66 respondents who stated that they had previously taken some version of the Moral Sense Test, and 25 more with apparently frivolous demographic responses (age <11 or >97, residence in Antarctica, or graduate degree obtained before age 20). We also excluded individual responses when the reading and response time was under 4 seconds (2% of responses).

### 2.2 Questionnaire Design

The survey consisted of several demographic questions (age, gender, education level, etc.), then 17 hypothetical scenarios, each requiring a moral judgment, followed by 5 questions about general moral principles. The Supplementary Online Material contains the full text.

*Double effect scenarios.* Questions 1 and 2 involved judgments about saving five lives at the expense of one. A Push-type scenario involved killing one person through direct physical contact as a means of saving five people. A Switch-type scenario involved one person's dying, without direct physical contact from the agent, as a side effect of an action to save five. There were four versions of each scenario type, differing in context: a runaway boxcar, a fire, a boat, and a hospital. Questions 1 and 2 comprised one Push and one Switch scenario drawn from the same scenario context, in random order. Respondents rated the hypothetical action on a seven-point scale from (1) 'extremely morally good' to (7) 'extremely morally bad' with the midpoint (4) labeled 'neither good nor bad'.

Questions 14-17 also included double effect scenarios. In addition to the Push and Switch scenarios, we presented two other types expected to receive intermediate responses but which are not targets of the present analysis. We presented the four scenario types in random order, always in a different scenario context than had appeared in Questions 1-2.

Questions 3-5 concerned killing one to save many (e.g., in an epidemic), were unvaried in order, and are also not targets of the present analysis.

*Action-omission scenarios.* Questions 6-9 involved judging actions versus omissions. The 'Vest' scenario pair involved snatching a life vest from a drowning person to increase one's own safety ('Take' Action) or failing to offer one's life vest ('Not Give' Omission). The 'Oxygen' scenario pair involved either taking away a troubled diver's oxygen line for one's own use ('Take') or failing to sacrifice one's own oxygen line ('Not Give'). Order of presentation was counterbalanced between participants, either AOOA or OAAO. Responses employed the same scale as in the double effect scenarios. Half of the respondents saw a version of the Vest scenarios in which the drowning victim is described in the second person, as 'you', rather than as 'a man'. Order effects were consistent across both phrasings, so we merged both types in the analyses below.

*Moral luck scenarios.* Questions 10-13 concerned moral luck. In one scenario pair, a drunk driver passes out and discovers either that he has hit a tree (Good Luck) or that he has killed a girl (Bad Luck). Another scenario pair involved a negligent construction worker either killing or not killing a pedestrian below. Order of presentation was counterbalanced between participants, either GBBG or BGGB. The order of the moral luck scenarios was yoked to the order of the action-omission scenarios so that GBBG was always paired with AOOA and BGGB was always paired with OAAO. Responses were on a 7-point scale from (1) 'not at all morally blameworthy' to (7) 'extremely morally blameworthy', with the midpoint labeled 'substantially morally blameworthy'.

*Endorsement of principles.* The test ended with several questions about abstract principles. Question 18 concerned moral luck:

> Suppose two people do the exact same thing, with the exact same frame of mind. Then, due entirely to matters of chance beyond their control, one of them produces a very bad outcome, but the other does not. Should they receive different amounts of punishment or the same amount of punishment?

Response options were 'same' or 'different'. Question 19 concerned the action-omission principle:

> Sometimes you can save several people by actively and purposefully killing one person whom you could have let live. Other times you can save several people by purposefully allowing one person to die whom you could have saved. Is the first action morally better, worse, or the same as the second action?

Response options were 'better' 'worse' or 'same'. Question 20 concerned the doctrine of the double effect:

> Sometimes it is necessary to use one person's death as a means to saving several more people—killing one helps you accomplish the goal of saving several. Other times one person's death is a side-effect of saving several more people—the goal of saving several unavoidably ends up killing one as a consequence. Is the first morally better, worse, or the same as the second?

Response options were 'better' 'worse' or 'same'. Questions 21 and 22 concerned the moral relevance of physical contact and general normative ethical stance (deontological, consequentialist, or virtue-based) and are not targets of the present analysis.

We recognize that much can turn on exactly how the above principles are stated; we aimed for simple statements comprehensible to non-specialists. For example, we recognize that the action-omission distinction may look different depending on whether one's motive is self-

interest or charity and that one might endorse moral luck concerning punishment but not blameworthiness. For this reason among others, endorsement or rejection of the various general principles is consistent with a variety of responses to the scenario types. Our analyses do not test for consistency between scenario ratings and principle endorsements, but rather test whether variation in presentation order of the scenarios affected endorsement of the general principles. We did find the expected associations between scenario judgments and endorsements of related principles (e.g., respondents who rated the Good Luck and Bad Luck scenarios inequivalently were more likely to endorse the principle of moral luck), but these patterns are not informative for present purposes because such relationships might reflect either a pattern of principled moral reasoning from the outset or a pattern of post-hoc rationalization of prior scenario judgments.

### 2.3 Analysis

Our main analysis asks, for each scenario pair, whether the participant gave the same numerical response to each scenario ('equivalent' judgment) or rated the scenarios differently ('inequivalent' judgment) in the direction predicted the relevant principle: double effect, action-omission principle, or moral luck. We excluded cases in which the participant rated the scenarios differently in the *non*-predicted direction (5% of double effect cases, 8% of action-omission cases, and 1% of moral luck cases). So, for example, if a respondent rated a Push and a Switch scenario both as 5's on our 7-point scale, she rated the two scenarios equivalently; if she rated Push 6 and Switch 5 (thus, Push worse, the predicted direction), she rated the scenarios inequivalently; and if she rated Switch 7 and Push 6 (thus, Switch worse), the pair is excluded from the equivalency analysis. Similarly, our analysis of the endorsement of principles excluded cases in which the participant indicated that harmful omissions were *worse* than harmful actions

(8%) or that harmful side-effects were *worse* than harmful means (7%), and also cases in which the participant's prior scenario judgments targeting the relevant principle had been excluded due to low reaction time.

We also examined order effects on mean scenario ratings. However, we emphasize the equivalency analysis for four reasons. (1) The equivalency analysis is less subject to scaling concerns due to participants' using early cases to 'anchor' key points on the scale. (2) The doctrine of the double effect, action-omission distinction, and principle of moral luck are principles that directly concern the equivalency or inequivalency of different actions rather than the goodness or badness of those actions, so our focus on equivalency matches the focus of the philosophical literature. (3) Equivalency of response is more comparable across scenarios types. (4) For some scenarios median response was at ceiling, problematizing parametric analysis of means. However, the overall results of our analysis are similar whether we examine equivalency or means.

## 3. Results

### 3.1 Double Effect Scenarios

In Questions 1-2, respondents were more likely to rate the Push and Switch scenarios equivalently when Push was presented before Switch (70% vs. 54%, $Z = 8.1$, $p < .001$). All three participant groups showed similar effect sizes (Figure 1a), and for each the effect was statistically significant (Table 1).

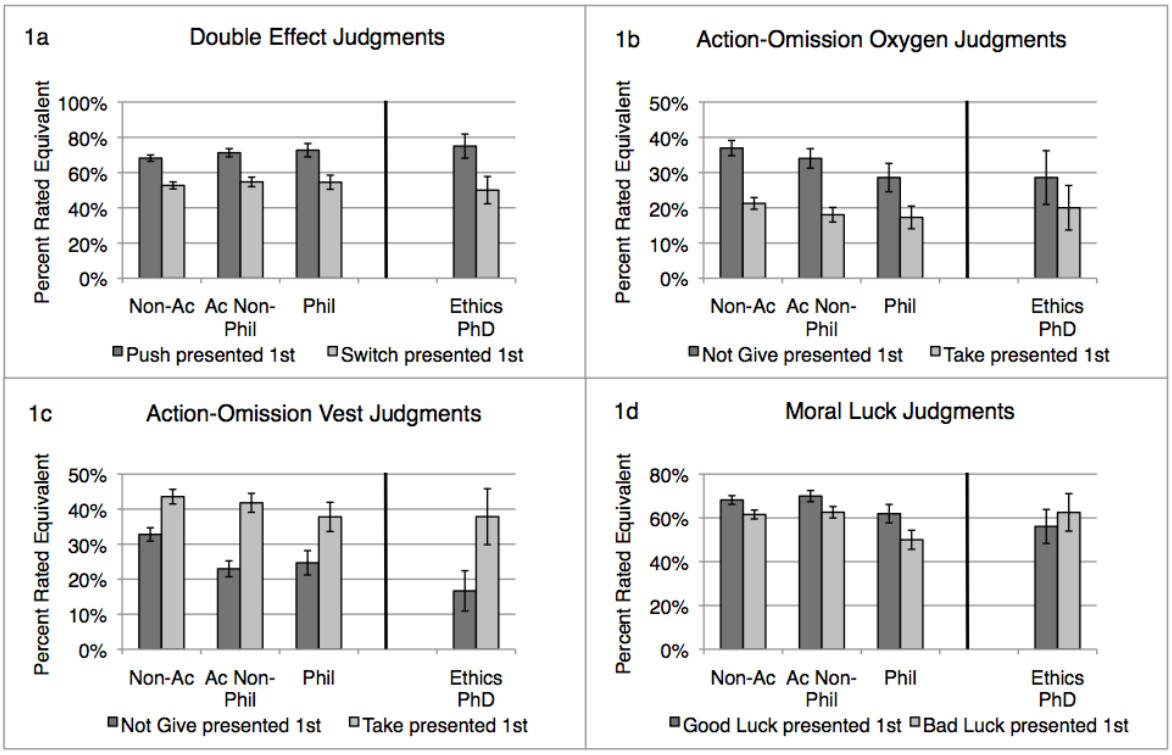


Figure 1: Effect of presentation order on percentage of scenarios rated equivalently across four types of scenarios: (a) double effect cases for Q1-Q2, (b) the action-omission oxygen case for Q6-10, (c) the action-omission vest case for Q6-10 and (d) moral luck cases for Q11-12.

| | Group | Percent rated equivalent: | 95% CI of difference in proportions | Z | p | Effect size h |
|---|---|---|---|---|---|---|
| | | **Switch First / Push First** | | | | |
| **Double Effect: Q1-2** | Non-Ac | 53% / 68% | 10%-21% | 5.8 | <.001 | 0.31 |
| | Ac Non-Phil | 55% / 71% | 10%-24% | 4.7 | <.001 | 0.33 |
| | Phil | 54% / 73% | 8%-29% | 3.4 | .001 | 0.40 |
| | Eth PhD | 50% / 75% | 0.5%-45% | 2.4 | .02 | 0.52 |
| | | **Switch First / Push First** | | | | |
| **Double Effect: Q14-17** | Non-Ac | 56% / 72% | 10%-21% | 5.4 | <.001 | 0.34 |
| | Ac Non-Phil | 62% / 74% | 5%-19% | 3.3 | .001 | 0.26 |
| | Phil | 68% / 74% | -6%-17% | 1.0 | .32 | 0.13 |
| | Eth PhD | 70% / 78% | -13%-29% | 0.8 | .45 | 0.18 |
| | | **Take First / Not Give First** | | | | |
| **Act vs. Omit: Oxygen** | Non-Ac | 21% / 37% | 10%-21% | 5.8 | <.001 | 0.36 |
| | Ac Non-Phil | 18% / 34% | 9%-23% | 4.6 | <.001 | 0.37 |
| | Phil | 17% / 29% | 1%-21% | 2.2 | .03 | 0.29 |
| | Eth PhD | 20% / 29% | -11%-28% | 0.9 | .39 | 0.21 |
| | | **Take First / Not Give First** | | | | |
| **Act vs. Omit: Vest** | Non-Ac | 44% / 33% | 5%-16% | 3.8 | <.001 | 0.23 |
| | Ac Non-Phil | 42% / 23% | 12%-26% | 5.4 | <.001 | 0.41 |
| | Phil | 38% / 25% | 2%-24% | 2.4 | .02 | 0.28 |
| | Eth PhD | 38% / 17% | 2%-40% | 2.2 | .03 | 0.48 |
| | | **Good Luck First / Bad Luck First** | | | | |
| **Moral Luck: Q10-11** | Non-Ac | 68% / 62% | 1%-12% | 2.3 | .02 | 0.13 |
| | Ac Non-Phil | 70% / 63% | 0.2%-15% | 2.0 | .04 | 0.15 |
| | Phil | 62% / 50% | 0.1%-24% | 2.0 | .048 | 0.24 |
| | Eth PhD | 56% / 63% | -29%-16% | -0.6 | .58 | -0.14 |
| | | **Good Luck First / Bad Luck First** | | | | |
| **Moral Luck: Q12-13** | Non-Ac | 61% / 66% | -11%-1% | -1.7 | .09 | -0.10 |
| | Ac Non-Phil | 67% / 67% | -8%-7% | -0.1 | .89 | -0.00 |
| | Phil | 50% / 59% | -21%-3% | -1.5 | .14 | -0.18 |
| | Eth PhD | 63% / 62% | -21%-25% | 0.2 | .88 | 0.02 |

Table 1: Effect of presentation order on percentage of scenarios rated equivalently for six types of scenarios, comparing across participant groups.

| Target Principle | Case | Mean when first | Mean when second | Difference (philosophers' difference) |
|---|---|---|---|---|
| Double Effect | Q1-2: Push | 4.37 | 4.13 | .24*** (.21) |
| | Q14-17: Push | 4.51 | 4.37 | .14 (.16) |
| | Q1-2: Switch | 3.38 | 3.88 | -.50*** (-.52**) |
| | Q14-17: Switch | 3.71 | 4.12 | -.41*** (-.35) |
| Act/Omit | Take Oxygen | 5.82 | 5.29 | .53*** (.39*) |
| | Not Give Oxygen | 4.53 | 4.02 | .51*** (.49***) |
| | Take Vest | 5.61 | 5.88 | -.27*** (-.20) |
| | Not Give Vest | 4.74 | 4.71 | .03 (-.05) |
| Moral Luck | Q10-11: Good Luck | 5.61 | 5.46 | .15* (.11) |
| | Q10-11: Bad Luck | 6.25 | 6.21 | .04 (.09) |

Table 2: Effect of presentation order on scenario mean ratings, all groups combined. T-test p values: * < .05, ** < .01, *** < .001.

Questions 14-17 showed order effects similar to those in Questions 1-2. When Push was presented before Switch among the four scenario types, 73% of respondents rated the two equivalently, versus 60% when Switch was presented before Push ($Z = 6.27$, $p < .001$). The order effect size was smaller for philosophers (Table 1), but not significantly (binary logistic regression, $Z = -1.3$, $p = .21$).

Analysis of means shows a similar pattern in order effect size among the groups (Table 2). Push was rated better when presented after Switch than when presented first, and Switch was rated worse when presented after Push than when presented first. Thus, respondents tended to assimilate their responses to the second scenario to their responses to the first scenario. However, Switch responses were considerably more labile than Push responses, explaining the higher rates of equivalency in scenario ratings when Push was presented first.

### 3.2 Action-Omission Scenarios

The Vest and Oxygen scenarios showed opposite order effects—the Vest cases showed greater equivalency for the action/omission order, while the Oxygen cases showed greater equivalency for the omission/action order. Thus, we analyzed order effects separately for the Vest and Oxygen cases. As with the double effect scenarios, the direction and magnitude of the order effects were similar among the groups (Figure 1b-c; Tables 1 and 2).

### 3.3 Moral Luck Scenarios

We also found order effects for the moral luck cases, and again these were comparable across the three major participant groups (Figure 1d; Tables 1 and 2), although absent for the ethics PhD subset.

Considering only the first presented scenario pair, participants were more likely to rate the Good Luck and Bad Luck cases equivalently when a Good Luck scenario was presented first (68% vs. 60%, $Z = 3.6$, $p < .001$). Ethics PhDs trended in the opposite direction (figure 1d and Table 1), but a binary logistic regression predicting equivalency from Good Luck-Bad Luck order and ethics PhD (vs. all others) found no significant interaction effect ($Z = 1.3$, $p = .21$), so it is not clear whether the lack of a similar effect among the ethics PhDs was due to chance.

The second Good Luck-Bad Luck scenario pair showed a marginally significant order effect in the opposite direction (62% vs. 66%, $Z = -1.9$, $p = .06$). Because scenarios were counterbalanced GBBG or BGGB, the observed reverse equivalency effect for the second pair may have reflected order effects carrying over from the first pair. For example, having judged the two drunk driver cases equivalently (hitting the tree vs. hitting the girl), participants may have been more likely to judge the subsequent pair of construction worker cases equivalently (killing a pedestrian vs. not killing a pedestrian), and vice versa for judgments of inequivalency.

Two-proportion analyses revealed no order effects on equivalency from one scenario type to any later scenario type (e.g., from double-effect order to act-omission judgments), nor between the order of presentation of the Push and Switch cases in Q1-Q2 and Push-Switch equivalency in Q14-Q17.

### 3.4 A Summary Measure of Order Effects on Equivalency Judgments

We aggregated equivalency order effects across all scenarios into a single summary statistic for each participant, facilitating an overall comparison of the magnitude of equivalency order effects between participant groups (Figure 2). The dependent variable was the number of scenario pairs (0 to 6) that the participant rated as equivalent. We included all scenario pairs analyzed above, including only participants whose data were included in all six analyses. The predictor was a variable indicating the number of scenarios (0 to 6) the participant viewed in the order favoring equivalency responses. These two variables were correlated ($r = .22$, $p < .001$), and to a similar extent for all groups: non-academics $r = .21$ ($p < .001$), academic non-philosophers $r = .19$ ($p < .001$), philosophers $r = .29$ ($p < .001$; ethics PhDs, $r = .35$, $p = .007$). Indeed, philosophers trended towards showing larger order effects than did the reference group of academic non-philosophers in a general linear model ($t = 1.8$, $p = .08$), including after controlling for age and gender ($t = 1.9$, $p = .07$).

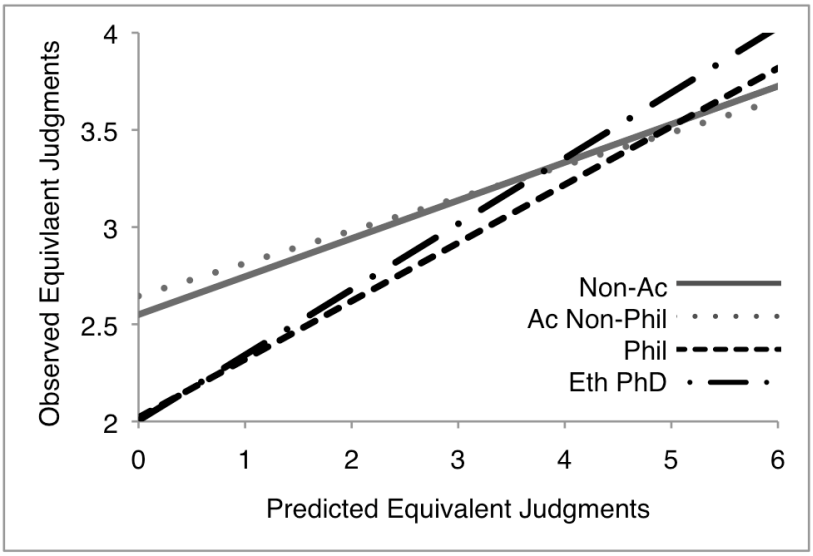


Figure 2: Linear regression trendlines for the number of scenario pairs rated as equivalent as a function of number of scenario pairs presented in an order favoring equivalent judgment. Greater slope indicates stronger equivalency order effects on moral judgment.

### 3.5 Order Effects on Endorsements of Moral Principles

The order of presentation of the scenarios showed little influence on non-philosophers' endorsements of moral principles. Non-philosophers who saw Bad Luck first were no more likely to endorse the principle of moral luck than those who saw Good Luck first (non-academics, 20% vs. 20%, $Z = 0.1$, $p = .95$; academic non-philosophers 18% vs. 18%, $Z = 0.3$, $p = .78$; Figure 3a). Those who saw Switch first were no more likely to endorse the doctrine of the double effect, and in fact less likely (non-academics, 46% vs. 53%, $Z = -2.2$, $p = .03$; academic non-philosophers, 51% vs. 55%, $Z = -1.1$, $p = .30$; Figure 3b). Given the scenario ratings results, the endorsement of the action-omission distinction should be favored when the first-presented scenario is Take Oxygen or Not Give Vest (as opposed to Not Give Oxygen or Take Vest). In

this case, there was a moderate effect on endorsement (non-academics, 52% vs. 44%, Z = 2.6, p = .01; non-philosopher academics, 59% vs. 53%, Z = 1.6, p = .11; Figure 3c).

In contrast, the order of presentation of the scenarios substantially influenced philosophers' subsequent endorsements of two of the three abstract moral principles. Philosophers were much more likely to endorse the principle of moral luck if they received a Bad Luck scenario first: 45% vs. 29% (Z = 2.7, p = .006); and they were also more likely to endorse the doctrine of the double effect if they saw Switch first: 62% vs. 46% (Z = 2.4, p = .02). However, philosophers who viewed Take Oxygen or Not Give Vest first were not more likely to endorse the action-omission distinction than the remainder who viewed Not Give Oxygen or Take Vest first, trending slightly in the opposite direction 54% vs. 58% (Z = -0.7, p = .49).

Turning to the ethics PhD subset of philosophers, for the doctrine of the double effect, endorsements differed by a degree similar to other philosophers: 59% vs. 40% (Z = 1.6, p = .12). However, this effect did not achieve statistical significance and must be interpreted cautiously due to the small sample size. As described above, ethics PhDs did not exhibit an equivalency order effect for their moral luck judgments, and therefore would not be predicted to exhibit a corresponding rationalization effect on endorsement. Indeed, order effects did not significantly influence ethics PhDs' endorsement of the principle of moral luck (38% vs. 38%, Z = 0.0, p = .96). Similarly to philosophers as a group, ethics PhDs were not significantly more likely to endorse the action-omission distinction when viewing Take Oxygen and Not Give Vest first (56%) than when viewing Not Give Oxygen and Take Vest first (56%; Z = 0.0, p = .99).

Across all participant groups, willingness to endorse one moral principle (e.g., moral luck) was correlated with willingness to endorse other moral principles (e.g., double effect). This was apparently due in part to a 'sequential endorsement effect': Endorsing one moral

principle earlier in a sequence made participants more likely to endorse another moral principle later in that sequence. Philosophers who viewed the Bad Luck case first, and were thus more likely to endorse moral luck, were significantly more likely than those who viewed the Good Luck case first to endorse the doctrine of double effect (64% vs. 44%, $Z = 3.2$, $p = .001$) and the action-omission distinction (63% vs. 49%, $Z = 2.1$, $p = .03$), endorsement choices that were presented *after* moral luck endorsements. There was no effect of double effect scenario order on moral luck or action/omission endorsements, endorsement choices presented *before* double-effect endorsements.

We took advantage of the sequential endorsement effect to measure the maximum influence of scenario order on endorsement by comparing philosophers' endorsement of the doctrine of the double effect among participants who viewed both the Switch case and Bad Luck case first to those who viewed both cases second. Among all philosophers, 70% who viewed both cases first endorsed the doctrine of the double effect, compared with 28% who viewed both cases second ($Z = 4.7$, $p < .001$). Among ethicists, 62% who viewed both cases first endorsed the doctrine of the double effect, compared with 28% who viewed both cases second ($Z = 2.0$, $p = .049$).

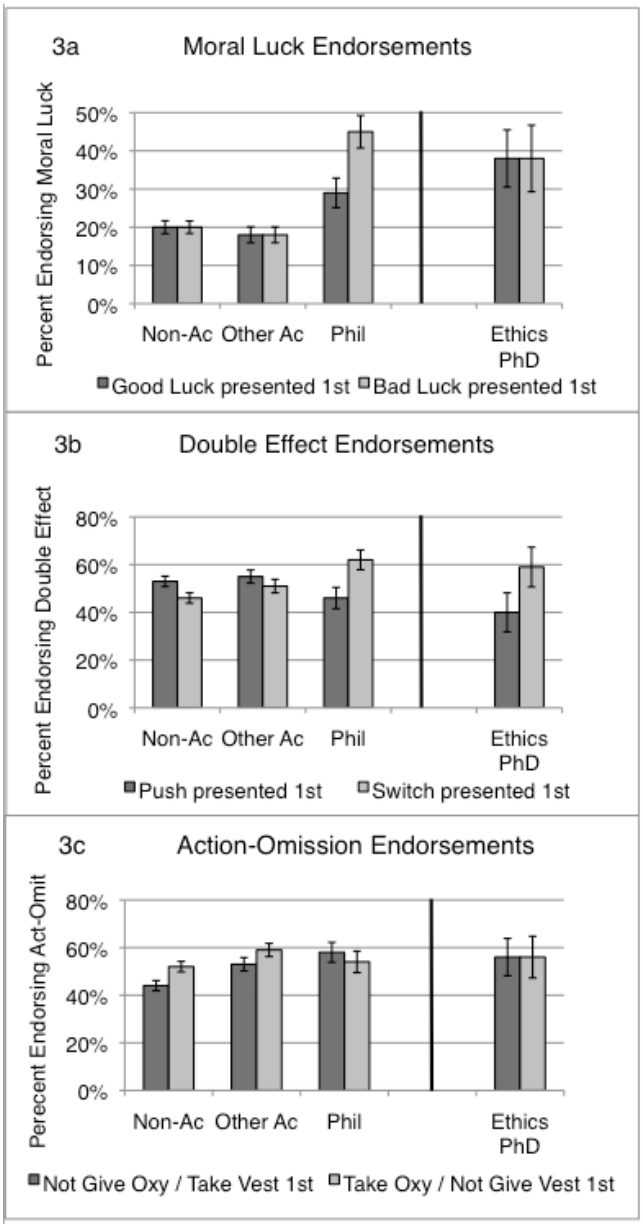


Figure 3: Percentage of participants endorsing (a) the principle of moral luck depending on the order of presentation of moral luck scenarios, (b) the doctrine of the double effect depending on the order of presentation of first double effect scenarios, and (c) the action-omission distinction depending on order of presentation of the action-omission scenarios.

### 3.6 A Summary Measure of Order Effects on Endorsement of Principles

We aggregated order effects on endorsement across all three principles into a single summary statistic for each participant in order to compare of the overall magnitude of those effects between participant groups (Figure 4). Our dependent variable was the number of principles (0 to 3) that the participant endorsed. We used the same predictor as in our aggregate analysis of equivalency order effects during scenario judgment: the number of scenarios (0 to 6) the participant viewed that had been presented in the order favoring equivalency responses. We predicted that participants who viewed cases in the order favoring equivalency would be less likely to endorse moral principles favoring inequivalency between the cases. As predicted, these two variables were negatively correlated, though only slightly ($r = -.07$, $p = .006$). This relationship was largest and statistically significant for philosophers ($r = -.17$, $p = .009$), with the largest effect size for ethics PhDs ($r = -.20$, $p = .11$). Smaller trends were evident but non-significant for non-academics ($r = -.05$, $p = .15$) and academic non-philosophers ($r = -.05$, $p = .23$). A comparison of the effect between philosophers and non-philosopher academics in a general linear model was marginally significant ($t = -1.7$, $p = .08$; and $t = -1.8$, $p = .07$ after controlling for gender and age).

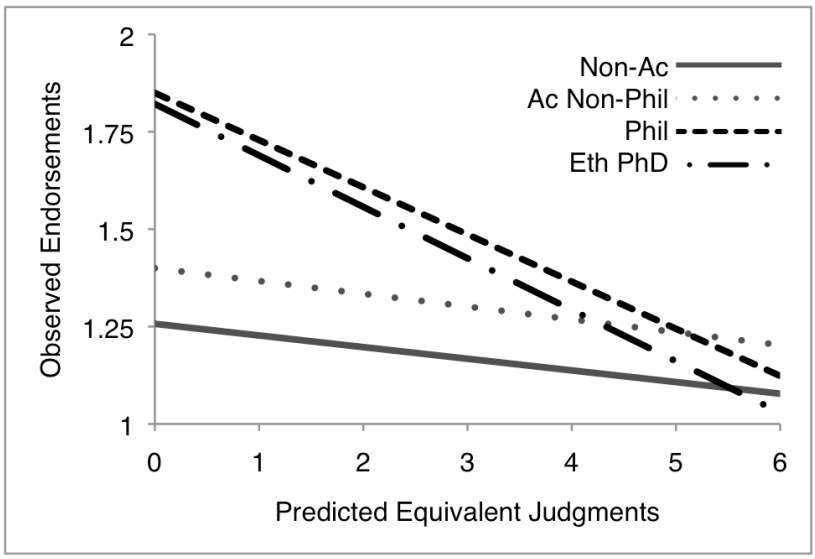


Figure 4: Linear regression trendlines for the number of moral principles endorsed that favor inequivalency as a function of number of scenario pairs presented in an order favoring equivalent judgment. Greater slope with negative sign indicates larger order effects on the endorsement of general principles.

## 4. Discussion

To the extent that judgments about individual scenarios are driven by stable moral principles, those judgments should not be affected by order of presentation of the scenarios. And to the extent that people choose to endorse or reject moral principles for stable, consistent reasons, those decisions should not be strongly influenced by the order in which several previous judgments were made. Philosophers—especially ethics PhDs at well-ranked research departments—should seemingly be particularly resistant to order effects on their scenario judgments and endorsements of principles, due to prior familiarity with the principles and general types of scenarios. However, even this 'best-case' group of participants showed

substantial order effects on their judgments about moral scenarios and their endorsement of moral principles. Our analysis found no support for the view that philosophical expertise enhances the stability of moral judgment against order effects; it suggests, instead, that philosophical expertise may actually enhance post-hoc rationalization.

### 4.1 Order Effects on Judgments about Particular Scenarios

Both philosophers and non-philosophers showed significant order effects for all three types of scenario. In our summary measure of order effects across all scenario judgments, philosophers and ethics PhDs trended marginally higher than the comparison groups. Thus, philosophers showed no greater tendency than non-philosophers to use the consistent application of moral principles to reduce order effects on their scenario judgments. Of course, there may be contexts in which philosophers will excel at the application of explicit principles, for example, in evaluating the validity of a proposed deductive syllogism. Philosophers may also more skillfully apply general moral principles to specific moral cases, but this is difficult to assess in a way that unambiguously distinguishes principled reasoning from post-hoc rationalization. It is precisely this methodological challenge that motivated our use of order effects as a metric for the consistent application of principled reasoning. It is particularly striking that philosophical expertise did not reduce order effects for cases intended to target the doctrine of the double effect, the action-omission distinction, and the principle of moral luck, given that these philosophical principles are widely discussed in terms of hypothetical scenario comparisons very much like those we presented to our participants.

Our experiment was not designed to clarify the psychological basis of the order effects on scenario judgments, and the order effects we observed were variable in size and direction. For

instance, our two pairs of action-omission cases produced opposite effects. However, it is likely that order effects between closely-matched pairs of hypothetical scenarios reflect a general desire to maintain consistency in judgment (see also Lombrozo, 2009). For example, having judged that it is morally bad to push a man in front of a train to save five others, some participants may resist the apparent inconsistency in judging that it is permissible to flip a switch that produces the same consequences. Accordingly, we suggest that order effects arise from an interaction between intuitive judgment and subsequent explicit reasoning: The intuition elicited by the first case becomes the basis for imposed consistency in the second case (Lombrozo, 2009). When the intuition elicited by one case is 'stronger'—that is, more resistant to revision by explicit reasoning—than the intuition elicited by the complementary scenario, this would lead to the asymmetric equivalency effects that we report here. When the stronger case comes first, it would exert a relatively larger influence on the subsequent judgment of the weaker case, making it more likely for the cases to be judged equivalently; but when the weaker case comes first, it would exert a lesser influence on the stronger case, leading to more inequivalent judgments. To take the familiar example of the trolley problem, it has been proposed that the 'push' version engages an automatic, affective response that the 'switch' case does not (Cushman, Young, & Greene, 2010; Greene, et al., 2001; Greene et al. 2009). This may explain why judgments of the switch case are apparently more malleable under the influence of prior push judgments, whereas push judgments are comparatively stable. The simplest interpretation of our findings is that this interaction between automatic processes and explicit reasoning, as well as the general desire to impose consistency between judgments, operates similarly in philosophical experts and novices.

An alternative explanation for the order effects we have identified is that certain cases presented new information or highlighted new considerations relevant to the judgment of the

other cases. For instance, the Push version of the trolley problem might highlight the rights of a single victim against harmful intervention, consequently exerting an influence on the judgment of the Switch version. Participants viewing the cases in different orders would then have access to identical information after, and only after, both cases have been presented. If this alternative explanation of the order effects is correct, order of presentation should not have any influence on the subsequent endorsement of a moral principle that distinguishes the cases: By the time of endorsement, all participants would have seen (for example) the rights-highlighting case. Yet we did find order effects on the endorsement of principles, and we turn next to consider this finding.

### 4.2 Order Effects on Endorsements of Abstract Moral Principles

Professional philosophers who viewed the first pair of double-effect or moral luck scenarios in an order favoring inequivalent judgment were more likely to subsequently endorse the doctrine of the double effect and the principle of moral luck—principles favoring inequivalent treatment of the scenarios. For these scenario types we observed no corresponding effect among non-philosophers. Conversely, for the action-omission principle we did not find the predicted effect of order on endorsement among philosophers, but we did identify a small effect for non-philosophers. Aggregating across all three principles we found a significant order effect on philosophers' endorsements of general moral principles that was three times larger than the corresponding, non-significant effect for non-philosophers.

Thus, it appears that a factor that we assume philosophers would deem irrelevant—order of presentation of cases—can exert a large influence on professional philosophers' judgments about abstract moral principles, presumably without their awareness. This effect is particularly striking because, regardless of the order of presentation, all philosophers had viewed and judged

the same pairs of cases by the time they were asked about the general principles. The effect sizes are also striking: For example, the joint effect of the order of presentation of the moral luck and double effect cases was to shift philosophers' rates of endorsement of the doctrine of the double effect from 28% to 70%, including 28% to 62% for ethics PhDs—a very large change considering how familiar and widely discussed the doctrine is within professional philosophy.

Rationalization, as we use the term, occurs when automatic, intuitive processes drive moral judgments and explicit moral reasoning is recruited only after the fact to justify those judgments, normally proceeding without introspective access to the original processes driving the judgments. Perhaps the simplest interpretation of our results is that philosophers' skill at moral reasoning is most effective during post-hoc rationalization. That is, philosophical expertise provides no protection against unwanted order-effect biases on moral judgments about particular scenarios, and philosophers' labile subsequent reasoning about abstract moral principles follows where their judgments about particular scenarios lead. However, even if we accept this interpretation, the magnitude of such rationalization remains to be determined. The effect sizes we report, though large, are consistent with the possibility that a majority of philosophers adhere consistently to principles.

It is notable that non-philosophers' endorsements of moral principles appear to be substantially less influenced by the order of presentation of the particular scenarios. We suggest two complementary hypotheses that would account for this result. First, non-philosophers might have lacked the conceptual resources necessary to recognize the relationship between their initial judgments and their subsequent endorsements of abstract principles. This explanation seems particularly likely for the doctrine of the double effect, which involves a non-obvious conceptual distinction between harm intended as a means and harm as a foreseen side effect. However, it

seems less likely for the principle of moral luck, which deals with the more familiar concepts of recklessness, accidents, and punishment. Second, philosophers might be more motivated to impose consistency between their judgments about specific cases and their endorsements of abstract principles. On the first explanation, philosophers are more *able* to rationalize; on the second, they are more *motivated* to rationalize.

## 5. Conclusion

The method of philosophy is often characterized as a matter of reconciling intuitive judgments about particular cases with plausible general principles (Bealer, 1998; Fischer & Ravizza, 1992; Rawls, 1971). While the psychological basis of ordinary people's judgments about particular moral cases may often be very different from the principles they invoke to rationalize those judgments (Carlsmith, Darley, & Robinson, 2002; Cushman, et al., 2006; Haidt, 2001; Hauser, et al., 2007), it has been unclear to what extent this is also true of professional philosophers. Our results suggest that even professional philosophers' judgments about familiar types of cases in their own field can be strongly and covertly influenced by psychological factors that they would not endorse upon reflection, and that such unwanted influences can in turn strongly influence the general principles those philosophers endorse.

Department of Philosophy

University of California at Riverside

Department of Psychology

References

Baron, J., & Ritov, I. 2004: Omission Bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, 94, 74-85.

Bartels, D. 2008: Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition,* 108, 381-417.

Bealer, G. 1998: Intuition and the autonomy of philosophy. In M. R. DePaul & W. Ramsey (Eds.), *Rethinking Intuition: The psychology of intuition and its role in philosophical inquiry*. Lanham, MD: Rowman & Littlefield.

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

Carlsmith, K., Darley, J., & Robinson, P. 2002: Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, 83, 284-299.

Crosthwaite, J. 1995: Moral expertise: A problem in the professional ethics of professional ethicists. *Bioethics,* 9, 361-379.

Cushman, F. A. 2008: Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108, 353-380.

Cushman, F. A., Dreber, A., Wang, Y., & Costa, J. 2009: Accidental outcomes guide punishment in a 'trembling hand' game. *PLOS One*, 4, e6699.doi:6610.1371/journal.pone.0006699.

Cushman, F. A., Young, L., & Greene, J. D. 2010: Multi-system moral psychology. In J. Doris et al. (Eds.), *The Oxford Handbook of Moral Psychology*: Oxford University Press.

Cushman, F. A., Young, L., & Hauser, M. D. 2006: The role of conscious reasoning and intuitions in moral judgment: Testing three principles of harm. *Psychological Science*, 17, 1082-1089.

Ditto, P. & Liu, B. in press: Deontological Dissonance and the Consequentialist Crutch. In M. Mikulincer & P. R. Shaver (Eds.), *Social psychology of morality: The origins of good and evil*: APA Press.

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

Føllesdal, A. 2004: The philosopher as coach. In E. Kurz-Milcke & G. Gigerenzer (Eds.), *Experts in Science and Society*. New York: Kluwer.

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

Frederick, S. 2005: Cognitive reflection and decision making. *Journal of Economic Perspectives,* 19 (4), 25-42.

Gino, F., Shu, L., & Bazerman, M. 2010: Nameless + harmless = blameless: When seemingly irrelevant factors influence judgment of (un) ethical behavior. *Organizational Behavior and Human Decision Processes,* 111, 93-101.

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

Haidt, J. 2001: The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814-834.

Haidt, J., & Hersh, M. A. 2001: Sexual morality: The cultures and emotions of conservatives and liberals. *Journal of Applied Social Psychology*, 31, 191-221.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. 2007: A dissociation between moral judgment and justification. *Mind & Language*, 22, 1-21.

Hofmann, F. 2010: Intuitions, concepts, and imagination. *Philosophical Psychology*, 23, 529-546.

Hovarth, J. 2010: How (not) to react to experimental philosophy. *Philosophical Psychology*, 23, 447-480.

Howard-Snyder, F. 2002/2007: Doing vs. allowing harm. *Stanford Encyclopedia of Philosophy* (Spring 2010 edition).

Kohlberg, L. 1984: *The Psychology of Moral Development.* Cambridge, MA: Harper & Row.

Kuhn, D. 1991: *The Skills of Argument*. Cambridge University Press.

Leaman, G. 1993: *Heidegger im Kontext*. Hamburg: Argument-Verlag.

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

Lombrozo, T. 2009: The role of moral commitments in moral judgment. *Cognitive Science, 33*, 273-286.

McIntyre, A. 2004/2009: Doctrine of double effect, *Stanford Encyclopedia of Philosophy* (Spring 2010 edition).

Mikhail, J. M. 2000: *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in A Theory of Justice'*. Ph.D. dissertation in Philosophy. Ithaca: Cornell University.

Mikhail, J. 2009: Moral grammar and intuitive jurisprudence:: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation*, 50, 27-100.

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

Nelkin, D. K. 2004/2008: Moral luck. *Stanford Encyclopedia of Philosophy* (Spring 2010 edition).

Paxton, J. M. and Greene, J. D. 2010: Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2, 511-527.

Petrinovich, L., & O'Neill, P. 1996: Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology*, 17, 145-171.

Pizarro, D. A., & Bloom, P. 2003: The intelligence of the moral intuitions: comment on Haidt (2001): *Psychological Review, 110*, 193-196.

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

Rawls, J. 1971: *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rest, J. R. 1993: Research on moral judgment in college students. In A. Garrod (Ed.), *Approaches to Moral Development*. New York: Teachers College.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. 2008: Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096-1109.

Schwitzgebel, E., & Rust, J. 2009: The moral behaviour of ethics professors: Peer opinion. *Mind, 118*, 1043-1059.

Schwitzgebel, E., & Rust, J. 2010: Do ethicists and political philosophers vote more often than other professors? *Review of Philosophy and Psychology,* 1, 189-199.

Schwitzgebel, E., & Rust, J. in preparation: The self-reported moral behavior of ethics professors.

Schwitzgebel, E., Rust, J., Moore, A., Huang, L., & Coates, J. in press: Ethicists' courtesy at philosophy conferences. *Philosophical Psychology.*

Singer, P. 1972: Famine, affluence, and morality. *Philosophy & Public Affairs, 1*, 229-243.

Spranca, M., Minsk, E., & Baron, J. 1991: Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27, 76-105.

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

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

Wheatley, T., & Haidt, J. 2005: Hypnotic disgust makes moral judgments more severe. *Psychological Science*, 16, 780-784.

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

Williamson, T. in press: Philosophical expertise and the burden of proof. *Metaphilosophy*.

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

Young, L., Nichols, S., & Saxe, R. 2010: Investigating the neural and cognitive basis of moral luck: It's not what you do but what you know. *Review of Philosophy and Psychology*, 1, 333-349.