

Gold, Natalie , Pulford, Briony D. & Colman, Andrew M. (2015) Do as I say, don't do as I do: differences in moral judgments do not translate into differences in decisions in real-life trolley problems. *Journal of Economic Psychology,* 47, 50 - 61. https://doi.org/10.1016/j.joep.2015.01.001

https://researchonline.lse.ac.uk/id/eprint/109257/

Version: Published Version
Licence: Creative Commons: Attribution 4.0

LSE Research Online is the repository for research produced by the London School of Economics and Political Science. For more information, please refer to our Policies page or contact lseresearchonline@lse.ac.uk

Journal of Economic Psychology 47 (2015) 50–61

Contents lists available at ScienceDirect

Journal of Economic Psychology

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

Do as I Say, Don't Do as I Do: Differences in moral judgments do not translate into differences in decisions in real-life trolley problems

Natalie Gold [a,*], Briony D. Pulford [b,1], Andrew M. Colman [b,1]

[a] Department of Philosophy, King's College London, UK
[b] School of Psychology, University of Leicester, UK

ARTICLE INFO

Article history:
Received 29 July 2014
Received in revised form 3 November 2014
Accepted 6 January 2015
Available online 12 January 2015

JEL classification:
D03
D63
D64
C91

PsycINFO classification:
2340
2630

Keywords:
Moral behavior
Moral decision making
Moral judgments
Trolley problems

ABSTRACT

Many people judge that it is permissible to harm one person in order to save many in some circumstances but not in others: it matters how the harm comes about. Researchers have used trolley problems to investigate this phenomenon, eliciting moral judgments or behavioral predictions about hypothetical scenarios where five people can be saved at the cost of harming one other person. We operationalized trolley problems in the laboratory, with economic incentives and real-life consequences, allowing us to observe not only judgments but actual decisions. We varied whether the five were saved by clicking a switch that diverted the harm to the one or by dragging the one in front of the harm. We found differences in moral judgments between the two tasks, but no differences in behavior. The judgments of actors and observers also differed, with observers judging it more right to act. Our results suggest that the difference between moral judgments and actions arises because participants think that doing the right action still involves doing something morally discreditable, and that the morality of taking action does not exhaust the normative reasons for acting.
© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Trolley problems were devised by philosophers in order to investigate why it is permissible to cause a harm to one person in order to save many in some circumstances but not in others (Foot, 1967; Thomson, 1976, 1985). The paradigm trolley problem is *Side-track*: There is a runaway trolley that threatens to kill five men on the track ahead. An agent can save the five by switching a lever that will divert the trolley onto a Side-track. However, on the Side-track is one man, who would be killed. This contrasts with *Footbridge*, where the agent can save the five by pushing a large man off a footbridge in front of the trolley, stopping the trolley but killing the one. In both cases, the decision is whether to take an action that results in

* Corresponding author at: Philosophy Department, King's College London, Strand, London WC2R 2LS, UK. Tel.: +44 (0)20 7848 2750.
E-mail addresses: natalie.gold@rocketmail.com (N. Gold), bdp5@le.ac.uk (B.D. Pulford), amc@le.ac.uk (A.M. Colman).
[1] Address: School of Psychology, University of Leicester, Leicester LE1 7RH, UK. Tel.: +44 (0)116 229 7197.

http://dx.doi.org/10.1016/j.joep.2015.01.001
0167-4870/© 2015 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

the death of one person in order to save five. However, many people have the intuition that it is morally permissible to turn the trolley in Side-track but not to push the man in Footbridge. It matters how the harm to the one and the saving of the five come about.

There is a lot of evidence that people make different moral judgments in hypothetical Side-track and Footbridge problems (Gold, Pulford, & Colman, 2013; Greene et al., 2009; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2011). However, there is little evidence about people's actual behavior. Some previous experiments on actions in moral dilemmas have asked participants to predict their own behavior (e.g. Bartels, 2008; Petrinovich & O'Neill, 1996; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Tassy, Oullier, Mancini, & Wicker, 2013). Predictions of behavior have been shown to be notoriously unreliable (Osberg & Shrauger, 1986; Vallone, Griffin, Lin, & Ross, 1990). They may be especially problematic in moral dilemmas, because people's predictions may be biased toward whatever response they think is more socially desirable, even if this would not be reflected in their actual behavior (Koritzky & Yechiam, 2010). The best way to discover what people would do is to observe their actions, but we cannot operationalize life and death trolley problems in the laboratory.

Gold et al. (2013) found, using hypothetical scenarios, that the difference in moral intuitions between Side-track and Footbridge is preserved when the outcomes are economic harms. This suggests that we can study trolley problems using the methodology of experimental economics, which has already been used to investigate moral behaviors such as altruism, fairness, trust, cooperation, and reciprocity (e.g. Andreoni, Brown, & Vesterlund, 2002; Andreoni & Miller, 2002; Berg, Dickhaut, & McCabe, 1995; Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999, 2006; Houser & Kurzban, 2002). In this article, we report a laboratory experiment, comparing judgments and behavior in real-life Side-track and Footbridge scenarios, where decisions resulted in actual small economic harms to one or to five. We investigate whether there are behavioral differences between different trolley problems, what the patterns of moral judgments are in real-life trolley problems, and whether behavior corresponds to moral judgments.

1.1. Behavior in trolley problems

As far as we know, there are no previous comparisons of behavior in real-life trolley problems, but there are some relevant precursors. Navarrete, McDonald, Mott, and Asher (2012) studied behavior in Side-track in a virtual reality environment and found that 90.5% of participants turned the trolley. This is similar to the percentage of people who judge turning the trolley to be morally permissible in hypothetical scenarios. However, despite the heightened level of realism, the virtual reality environment still does not lead to actual outcomes for real people.

Côté, Piff, and Willer (2013) ran what was effectively a hypothetical version of Footbridge with small economic harms, asking participants whether they would take money from one other participant to benefit three others. This set-up is analogous to Footbridge for the following reason: An important difference between Side-track and Footbridge, according to moral philosophers, is that in Side-track the harm to the one occurs as a foreseen side effect when saving the five, whereas in Footbridge harming the one is a necessary means to saving the five. (If the one person was not there, then turning the train would still save the five in Side-track but the five in Footbridge could not be saved.) In Côté, Piff, and Willer's study, the one is harmed as a means to benefit the three, hence it is analogous to Footbridge but with small economic harms. They found that 60% of their participants said that they would take the money. However, the decisions were only hypothetical; the outcomes were not actually realized.

Our experimental design is based on a task used by Hsu, Anen, and Quartz (2008) who, while investigating the brain regions involved in trade-offs between equity and efficiency, used a computer animation where participants could flip a switch to divert a threat from one group of children to another. The outcomes were actually implemented, using small economic incentives. Gold, Colman, and Pulford (2014) adapted this methodology to study cultural variations in Side-track, finding that Chinese participants were less likely to divert the threat than British.

None of the preceding experiments compared behavior in Side-track and Footbridge. In this paper, we use the same methodology as Gold, Colman et al. (2014), supplemented with a Footbridge variant, to compare behavior and moral judgments in Side-track and Footbridge scenarios with real economic outcomes.

1.2. Relationship between behavior and moral judgments

Another reason to implement a real-life trolley task is to elicit moral judgments about a real situation, and to compare judgments with behavior. It is not obvious that moral judgments made in hypothetical situations will be the same as moral judgments made in real-life situations with actual consequences (FeldmanHall et al., 2012; Gold, Pulford, & Colman, 2014), or that behavior in trolley problems will correspond to moral judgments. People's actual moral behavior often does not live up to the moral attitudes that they express, not even that of ethicists (Schwitzgebel & Rust, 2014). People's predictions of their behavior in moral dilemmas are more utilitarian than their moral judgments, so that they are more likely to predict that they would harm one to benefit many than they are to judge that it is acceptable to do so (Kurzban, DeScioli, & Fein, 2012; Tassy et al., 2013).

Tassy et al. (2013) suggest three possible explanations for the divergence between moral judgment and predicted behavior. One possibility is that moral judgments and behavior are the outputs of separate psychological processes (Separate Processes Hypothesis). A second possibility is that differences between judgment and choice are due to *akrasia*, or

people acting against their own best judgments (Akrasia Hypothesis). A third is that judgments and behavior are made from different perspectives (Differing Perspectives Hypothesis). Drawing on a distinction made by Frith and de Vignemont (2005), Tassy et al. suggest that decisions about behavior are made from an "egocentric" perspective, which represents the situation relative to the decision-maker and, hence, emphasizes the consequences for the decision-maker, whereas judgments are made from an "allocentric" perspective, which represents the situation independent of the agent's own relationship to it.

These three different explanations for the discrepancy are potentially compatible, but they are not necessarily connected and may not all be responsible for the differences between judgments and actions. The data that we collect bears on the three hypotheses, as explained in Section 2.

2. Experimental design

2.1. Overview

We used a 2 (Scenario: Footbridge versus Side-track, within-subjects) $\times$ 2 (Role: Actor versus Observer, between-subjects) design. Participants were randomly allocated to be either actors or observers; their roles remained fixed during the experiment. Actors made decisions that influenced the amount of money donated to an orphanage in northern Uganda (following a similar protocol to Hsu et al., 2008), and observers were asked about what decision another person in the room should make. Our computer software recorded the time at which actors took action, if they did so, and the time taken for observers to say what the actor should do (their "action judgment"). Then both actors and observers rated the moral rightness of the action. All participants responded to both scenarios with the ordering counterbalanced to check for order effects.

Hence, as well as comparing moral judgments and behavior in Side-track and Footbridge, we are also able to compare actors' behavior with observers' overall judgments of what should be done (relevant to the Akrasia Hypothesis), the moral judgments and behavior of actors and of observers in the two scenarios (relevant to the Differing Perspectives Hypothesis), and response times (relevant to the Separate Processes Hypothesis).

2.1.1. Akrasia hypothesis

Akrasia occurs when people act against their *all-things-considered* judgment, or their overall judgment about what they should do. Experiments on moral decision making usually ask whether acting is morally acceptable or morally permissible. However, there may be conflicting moral requirements that apply to a situation. Philosophers argue about whether it must always be possible to weigh conflicting requirements and conclude that one overrides the other, or whether there are cases where neither requirement overrides the other and we have what philosophers call a "moral dilemma" (Sinnott-Armstrong, 1988). (Thus, for philosophers, "moral dilemma" is a technical term, whose meaning differs from the everyday meaning used in the psychology literature and in this paper.) In any case, asking about the morality of one possible action is not the same as eliciting an overall judgment about what should be done, and it is the comparison of behavior and overall judgments that is required to investigate the occurrence of akrasia. Asking actors for their all-things-considered judgments risks simply eliciting judgments that are consistent with their behavior. Hence we elicit the all-things-considered action judgments of observers, who cannot influence the actor's decision. The Akrasia Hypothesis implies that actors' decisions will differ from observers' judgments about what actors should do.

2.1.2. Differing perspectives hypothesis

The egocentric perspective is that of the person making the judgment or decision, whilst the allocentric perspective is an overview of the situation. Most trolley experiments are done from an allocentric point of view: the participant reads a narrative about the scenario and is asked to judge an action that may be performed by one of the agents within. Since we have both actors and observers, we can compare moral judgments and decisions made from an egocentric and an allocentric point of view. If judgments are allocentric, then we would expect the moral judgments of actors and observers to be the same; and if choices are egocentric, then behavior of actors and observers' action judgments may differ.

2.1.3. Separate processes hypothesis

Our computer software recorded the time taken for observers to make their action judgments, which allows us to test a dual process theory for judgments about moral behavior. Greene, Sommerville, Nystrom, Darley, and Cohen (2001) and Greene, Nystrom, Engell, Darley, and Cohen (2004) argue that we have an intuitive, emotional response to Footbridge, which drives moral disapproval of pushing the man in front of the train, but which can be over-ridden by moral reasoning. This dual process theory of moral judgment predicts that people who judge that acting is morally permissible in Footbridge will have a slower reaction time than those who find it immoral—because of the extra processing time needed to make a reasoned response rather than just to respond intuitively—but that there will be no difference in reaction time in scenarios such as Side-track that do not elicit such a strong emotional response. We can use our reaction time data to test whether the interaction effect predicted by dual process theory occurs for judgments about behavior. If it does not occur, then that will be consistent with the Separate Processing Hypothesis.

2.2. Participants

There were 176 participants: 69 men and 107 women, aged between 18 and 37 years (M = 20.91, SD = 3.50). They were all British and were recruited at the University of Leicester. Participants were either given course credits or paid £5 ($8) for their participation. They were tested in groups of 15–20.

2.3. Procedure and materials

Participants read a consent form and were assured of the anonymity of their data. After granting consent, they read a brochure from the Canaan Children's Home, depicting the children's plight and containing short biographies and photos of twelve of the orphans (who were matched for age and gender).

Participants were told that we had endowed each of the children in the photos with a sum of money enough to buy them one meal, and that each child would appear only once in the experiment. That amount was 30p (50c), although we did not tell the participants this. We expressed the decision in terms of meals because of the vastly greater purchasing power of money in Uganda than in the UK.

Participants then viewed an animation of a ball moving slowly across the computer screen towards a group of five children, represented by their photos (see screen shot in Fig. 1). There was also a photo of a single child on screen, who was not in the path of the ball. Participants were told that any children whose photo was hit by the ball would lose their meals. In the Side-track condition, actors had the option to click on a switch that flipped a lever, causing the ball to change direction and head towards the single child. In the Footbridge condition, actors had to use the mouse to drag the photo of the single child into the path of the ball that was heading towards the group of five, because the switch button and lever were absent. They had 11 s until the ball crossed the dotted line in the middle of the screen, in which they could click the switch or drag the photo. During this time, observers were asked whether the actors should click/drag (Yes/No) and to rate the statement that the actor should click/drag ("click" in the Side-track scenario and "drag" in Footbridge). For those actors who clicked the switch and for all observers, the time taken until their decision was recorded.

Before participants made their decisions and judgments, they watched two demonstrations of the animation, one where the lever was flipped or the photo was dragged, and one where it was not. In the demonstration there were no pictures, only blank rectangles, and the number of rectangles in each group was always five. The full animation took 17.5 s.

After the animation, participants were also asked to give a rightness rating: How wrong or right was it to flip the lever/drag the photo? Answers were on a 9-point scale (1 *Definitely wrong*, 5 *Neutral*, 9 *Definitely right*).

After completing their first scenario ("First Task"), the process was repeated, from the demonstration animations onwards, for the other scenario ("Second Task").

3. Results

3.1. Descriptive statistics

Some basic descriptive statistics are given in Tables 1 and 2, showing the proportion of each group who took the action (if they were actors) or said that the actor should take the action (if they were observers), and the mean rightness ratings of the action. The distribution of rightness judgments was bimodal, with a major peak on the "right" side and a minor peak on the "wrong" side, so we also show the mean rightness ratings for the groups split by those participants who took the action or

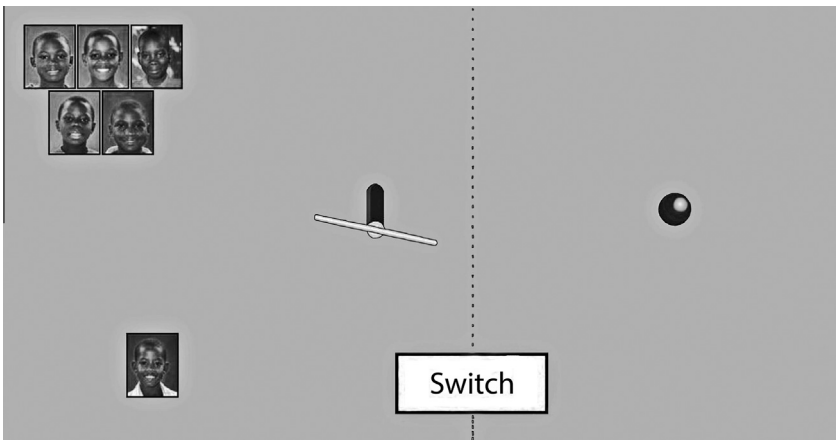

Fig. 1. Screenshot from the Side-track condition of our experiment.

Table 1
Descriptive statistics, First Task (first scenario completed by each participant).

	Actors				Observers			
	Side-track		Footbridge		Side-track		Footbridge	
Acted (actors)/judged the actor should act (observers)	80.0%		90.9%		72.1%		81.0%	
Right mean (*SD*)	5.78 (2.62)		4.91 (2.43)		6.41 (2.40)		6.26 (2.36)	
N	45		44		44		43	
Took action (actors)/judgment about action (observers)	No action	Action	No action	Action	Should not act	Should act	Should not act	Should act
Right mean (*SD*)	3.67 (3.08)	6.31 (2.24)	3.25 (2.50)	5.08 (2.39)	4.33 (3.09)	7.16 (1.53)	2.75 (0.89)	7.00 (1.78)
Time taken (s)	n/a	5.70	n/a	4.62	7.29	6.39	7.71	6.18
Mean (*SD*)	n/a	(2.34)	n/a	(2.46)	(1.65)	(1.65)	(2.22)	(1.86)
N	9	36	4	40	12	31	8	34

Table 2
Descriptive statistics, Second Task (second scenario completed by each participant).

	Actors				Observers			
	Side-track		Footbridge		Side-track		Footbridge	
Acted (actors)/judged the actor should act (observers)	95.5%		88.9%		92.9%		90.9%	
Right mean (*SD*)	5.41 (2.31)		5.51 (2.43)		7.02 (1.70)		7.07 (1.56)	
N	44		45		43		44	
Took action (actors)/judgment about action (observers)	No action	Action	No action	Action	Should not act	Should act	Should not act	Should act
Right mean (*SD*)	4.50 (3.54)	5.45 (2.29)	3.20 (1.79)	5.80 (2.36)	4.00 (1.73)	7.33 (1.40)	4.00 (2.58)	7.37 (1.06)
Time taken (s)	n/a	4.66	n/a	3.37	7.45	5.42	6.57	5.34
Mean (*SD*)	n/a	(2.33)	n/a	(2.27)	(2.27)	(1.92)	(2.55)	(2.26)
N	2	42	5	40	3	39	4	40

said the actor should take the action, and those participants who did not take the action or said that the actor should not take the action.

There were order effects for behavior in the Side-track task, for both actors and observers. Actors who did Side-track second were more likely to click the switch, $\chi^2(1, 89) = 4.91$, $p = .027$, effect size $\varphi = .235$ (small to medium effect), and observers who saw Side-track second were more likely to judge that the actor should click the switch, $\chi^2(1, 85) = 6.30$, $p = .012$, $\varphi = .272$ (small to medium effect). There were no order effects for behavior in footbridge.

There was also a trend for observers who saw footbridge second to judge dragging the photo as more right than those who saw it first, 7.07 vs. 6.26, $t(72.6) = 1.89$, $p = .063$, Cohen's $d = .44$ (small to medium effect). Otherwise, there were no order effects in rightness judgments. Because of these order effects, we analyze data from the first and second tasks separately.

3.2. First task

Surprisingly, those in the Side-track condition neither took action (actors) nor said that the actor should take action (observers) more often than those in the Footbridge condition. If anything, there was a marginally significant tendency for those in the Footbridge condition to take action, or to say that the actor should take action, more often than those in Side-track, $\chi^2(1, 174) = 2.28$, $p = .095$. There was no significant difference between the percentage of actors who took the action and the percentage of observers who said that they should, $\chi^2(1, 174) = 2.25$, $p = .133$.

A regression analysis of the rightness judgments reveals that, as expected, those who took the action (or said it should be taken) gave significantly higher rightness judgments than those who did not. In addition to this effect, observers gave significantly higher ratings than actors and those in the footbridge task gave lower rightness ratings than those in Side-track. See Model 1 in Table 3. There were no significant interaction effects, see Models 2–4 in Table 3, and the effects persist when we control for demographic effects, although we also find that older participants give higher ratings (see Model 5 in Table 3).

Many of our participants judged the action to be morally wrong even though they had taken the action or said that the actor should take it. We can group our participants according to whether their rightness ratings indicated that the action was wrong (0–4), neutral (5), or right (6–9). If the action was either neutral or right, but not morally wrong, then call it "permissible" (see Table 4). We can compare permissibility judgments with behavioral action or action judgments using a McNemar test, which shows that there were significant differences between actions or action judgments and permissibility judgments,

Table 3
Regression models of rightness judgments in the first task with dummy variables for actor/observer (observer = 1, actor = 0), side-track/footbridge (footbridge = 1, side-track = 0), and whether the participant took the action/said that the actor should take the action (yes = 1, no = 0), male/female (male = 1, female = 0). Standard errors in brackets beneath each coefficient.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	3.21***	3.39***	3.36***	3.79***	0.70
	(0.44)	(0.47)	(0.51)	(0.61)	(1.09)
Observer	1.20***	0.83*	1.21***	0.24	1.20***
	(0.33)	(0.46)	(0.33)	(0.77)	(0.33)
Footbridge	−0.83**	−1.20**	−1.25	−0.81**	−0.79**
	(0.33)	(0.46)	(0.79)	(0.33)	(0.33)
TakeAction	2.98***	2.99***	2.79***	2.30***	3.05***
	(0.43)	(0.43)	(0.55)	(0.65)	(0.42)
Observer ∗ Footbridge		0.75			
		(0.66)			
Footbridge ∗ TakeAction			0.51		
			(0.87)		
Observer ∗ TakeAction				1.18	
				(0.85)	
Age					0.11**
					(0.05)
Male					0.18
					(0.35)
R^2	0.26	0.27	0.26	0.27	0.29
Adjusted R^2	0.25	0.25	0.24	0.25	0.27

* Significant at the 10-percent level.
** Significant at the 5-percent level.
*** Significant at the 1-percent level.

Table 4
Distribution of right-neutral-wrong judgments and permissibility.

Condition	Rightness judgment			
	Wrong	Neutral		Right
Side-track actor				
Count	13	2		30
%	28.9%	4.4%		66.7%
Permissible			71.1%	
Side-track observer				
Count	7	2		35
%	15.9%	4.5%		79.5%
Permissible			84.1%	
Footbridge actor				
Count	18	5		21
%	40.9%	11.40%		47.7%
Permissible			59.1%	
Footbridge observer				
Count	10	1		32
%	23.3%	2.3%		74.4%
Permissible			76.7%	

$\chi^2(1, 174) = 5.60$, $p = .018$, $\varphi = .18$ (small to medium effect size). Usually, we would expect people to act only if they think that acting is permissible and hence that the number of participants who act (or say that action should be taken) would be less than or equal to the number of people who judge acting permissible. However, in three out of four conditions it was the other way round: more participants took action (or judged that action should be taken) than judged the action permissible (see Fig. 2). Many participants judged the action to be wrong but nevertheless took action (or judged that action should be taken), especially among the actors in the footbridge condition.

A two-way ANOVA on the time taken for observers to make their judgment of whether the actor should take action shows that there was no interaction effect between judgment (yes or no) and scenario, $F(1, 81) = 0.462$, $p = 0.499$. Observers who judged that the actor should click or drag made their judgments faster than those who judged that that actor should not act, $F(1, 81) = 6.86$, $p = .011$, $\eta^2 = .078$ (small to medium effect). There was no significant difference in timing between Side-track and Footbridge, $F(1, 81) = .047$, $p = 0.829$.

Actors acted faster than observers made their judgments. We only have response time data for those actors who clicked or dragged; the software timed out at 11s for those who did not act. But the actors who took action did so faster than observers made the judgment that they should act, $M = 5.13$ vs. $M = 6.28$, $t(134.9) = -3.25$, $p = 0.001$, $d = .56$ (medium effect). We

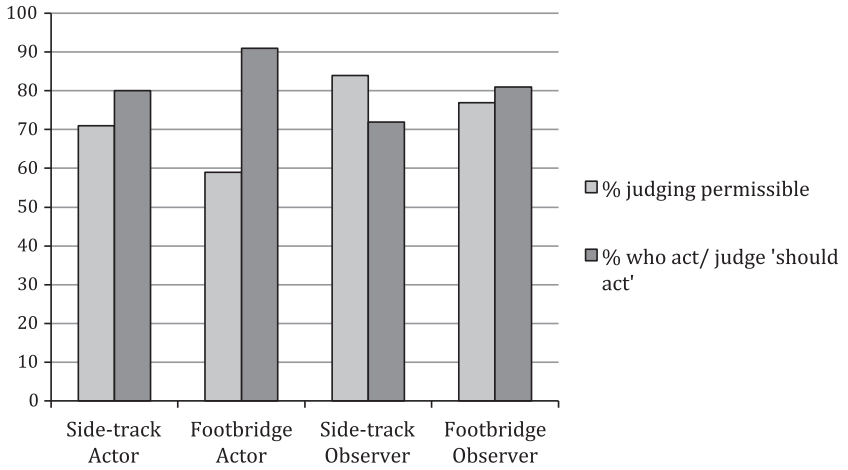

Fig. 2. Percentage who judged that acting was permissible in each condition of Task 1 compared to the percentage who acted/judged that the actor should act.

showed above that observers who judged "yes" the actor should act were faster than observers who judged "no", so actors were actually faster than all observers.

3.3. Second task

In the second task, there were no differences between groups in their propensity to take action or to say that the actor should take action. As we saw above, the number of participants taking action or saying that the actor should take action increased in Side-track. It seems that the numbers taking action in Footbridge in the first task were at ceiling, and, in the second task, the numbers taking action in Side-track increased so that they are also at ceiling.

A regression analysis shows that the actor-observer difference in rightness judgments persisted, with observers giving higher ratings than actors, but there is no longer a difference in ratings between Side-track and Footbridge. There are no inter-action effects and no demographic effects—the age effect we found in the first task has disappeared—and adding extra variables does not even improve the explanatory power of the model (see Table 5). The disappearance of differences between Side-track and Footbridge is consistent with there being order effects. Having completed the first task affected participants' views about acting, and the rightness of acting, in the second task.

Table 5
Regression models of rightness judgments in the second task with dummy variables for actor/observer (observer = 1, actor = 0), side-track/footbridge (footbridge = 1, side-track = 0), and whether the participant took the action/said that the actor should take the action (yes = 1, no = 0), male/female (male = 1, female = 0). Standard errors in brackets beneath each coefficient.

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5
Constant	2.87***	2.80***	3.47***	3.23***	2.27**
	(0.56)	(0.58)	(0.74)	(0.86)	(1.00)
Observer	1.63***	1.76***	0.45	1.62***	1.63***
	(0.29)	(0.41)	(1.01)	(0.29)	(0.29)
Footbridge	0.16	0.28	0.14	−0.39	0.13
	(0.29)	(0.40)	(0.29)	(1.06)	(0.29)
TakeAction	2.72***	2.73***	2.08***	2.35***	2.73***
	(0.53)	(0.53)	(0.75)	(0.87)	(0.53)
Observer ∗ footbridge		−0.26			
		(0.57)			
Observer ∗ action			1.28		
			(1.05)		
Footbridge ∗ action				0.59	
				(1.10)	
Age					0.02
					(0.04)
Male					0.29
					(0.31)
R^2	0.26	0.26	0.26	0.26	0.26
Adjusted R^2	0.24	0.24	0.25	0.24	0.24

* Significant at the 10-percent level.
** Significant at the 5-percent level.
*** Significant at the 1-percent level.

Table 6
Distribution of right-neutral-wrong judgments and permissibility in the second task.

Condition	Rightness judgment			
	Wrong	Neutral		Right
Side-track actor				
Count	12	10		22
%	27.3%	22.7%		50.0%
Permissible			72.7%	
Side-track observer				
Count	4	1		38
%	9.3%	2.3%		88.4%
Permissible			90.7%	
Footbridge actor				
Count	13	6		26
%	28.9%	13.3%		57.8%
Permissible			71.1%	
Footbridge observer				
Count	2	2		40
%	4.5%	4.5%		90.9%
Permissible			95.5%	

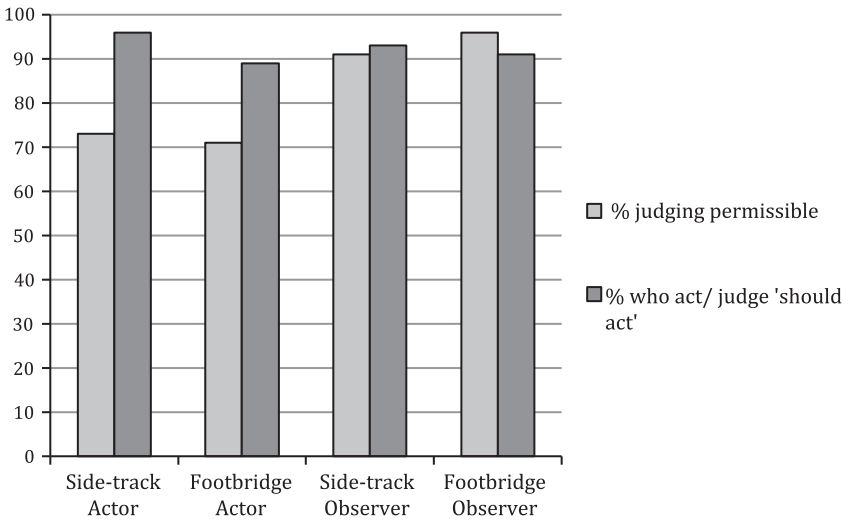

Fig. 3. Percentage who judged that acting was permissible in each condition of Task 2 compared to the percentage who acted/judged that the actor should act.

As in the first task, many of our participants judged the action to be morally wrong, even though they had taken the action or said that the actor should take it. Again, we can group our participants according to whether their rightness ratings indicated that the action was wrong (0–4), neutral (5), or right (6–9) and if the action was either neutral or right, but not morally wrong, then we call it "permissible" (see Table 6). We found that there were significant differences between actions/action judgments and permissibility judgments, using a McNemar test, $\chi^2(1, 175) = 8.04$, $p = .005$, $w = .21$ (small to medium). In this task, the effect seems to be driven by the actors, in both footbridge and Side-track, who took the action despite later rating it as wrong (see Fig. 3).

Participants reacted significantly faster in the second task than in the first, ($M = 4.79$s vs. $M = 5.88$s), as shown by a paired sample t-test, $t(158) = 5.72$, $p < .001$, Cohen's $d = .45$ (medium effect size). A two-way ANOVA on the time taken for observers to make their action judgment shows that there was no interaction effect between judgment (yes or no) and scenario, $F(1, 82) = .23$, $p = 0.635$. Observers who judged that the actor should click or drag made their judgments marginally faster than those who judged that that actor should not act, $F(1, 82) = 3.71$, $p = 0.058$, $\eta^2 = .043$ (small effect). There was no significant difference in timing between Side-track and Footbridge, $F(1, 82) = 0.32$, $p = 0.571$.

Again, actors acted faster that observers make their judgments, comparing the actors with the observers who judged that they should act (who we saw above were marginally faster than those who judged that the actor should not act), $M = 4.03$ vs. $M = 5.38$, $t(159) = -3.83$, $p < 0.001$, $d = .61$ (medium to large effect).

4. Manipulation check

Our Side-track and Footbridge scenarios preserve the main philosophical difference between Side-track and Footbridge. In Side-track the five are saved by diverting the ball and it is an unfortunate, but foreseen, side-effect that the one child's photo will be hit instead; in Footbridge, dragging the one child's photo into the path of the ball is a necessary means to stop the ball and save the five. Our scenarios also preserve the difference between pushing a button that moves a lever and using muscular force to move the one into the path, with our Footbridge scenario arguably involving the application of personal force, which Greene et al. (2009) argue influences moral judgments in trolley problems.

However, there is also evidence that Footbridge-style dilemmas elicit greater activity in brain regions involved with emotional processing than Side-track-style dilemmas (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Schaich Borg et al., 2006). Greene (2014) has argued that our automatic emotional response to Footbridge affects our moral judgments. It is not clear whether our real-life scenarios will have preserved the difference in emotional arousal between Side-track and Footbridge. On the one hand, there is real action involved, which should heighten arousal. But, on the other hand, the stakes are small and dragging a child's picture may not be as emotionally arousing as pushing someone off a footbridge. Therefore we ran a second experiment as a manipulation check, to test whether there is a difference in emotional arousal between our Side-track and Footbridge scenarios.

4.1. Method

4.1.1. Participants

There were 33 participants, 11 men and 22 women, aged between 18 and 28 (M = 19.64, SD = 2.23. They did the experiment as a course requirement. None of the participants had done our main experiment.

4.1.2. Procedure and materials

The participants were presented with the descriptions of two similar scenarios (within subjects design), one in which the action was "Dragging the photo of the child" (Footbridge) and the other was "Clicking the switch to move the lever" (Side-track). The scenarios were precise descriptions of the experiment described in Section 2.3, which the main study participants had taken part in. The presentation order of the two scenarios was counterbalanced. At the end of the main scenario they were given the options "Consider your options: OPTION A – Clicking the switch to move the lever, resulting in the five children keeping their meals and the one child losing his meal. OPTION B – Not intervening, resulting in the five children losing their meals and the one child keeping his meal."

After each scenario they were asked (in randomly presented order)

1. Which do you feel worse about doing? (1 *Much more option A* to 7 *Much more option B*, with 4 *No difference*).
2. Which arouses more emotion? (1 *Much more option A* to 7 *Much more option B*, with 4 *No difference*).
3. Which is more upsetting? (1 *Much more option A* to 7 *Much more option B*, with 4 *No difference*).

After they had completed both scenarios a final screen asked them to compare the two scenarios:

4. Thinking back to both scenarios you have just read about, if you were actually forced to take action, which of the two actions (dragging the photo vs. clicking the switch) would you feel worse about doing? (1 *Dragging the photo* to 7 *Clicking the switch*, with 4 *Equal*).

4.2. Results

In the direct comparison between the two tasks, participants overwhelmingly judged that they would feel worse about dragging the photo than clicking the switch. A one sample t-test comparing the results to a neutral score of 4 showed that the mean rating of 2.18 (SD = 1.31) was significantly different from the neutral point of 4, $t(32) = -7.973$, $p < .001$, $d = 1.389$ (very

Table 7
Means and standard deviations in ratings task (1 = clicking/dragging, 7 = not intervening).

Question	Scenario	
	Side-track	Footbridge
Which feels worse?	M = 3.82	M = 3.70
	SD = 2.02	SD = 2.08
Which arouses more emotion?	M = 4.00	M = 3.79
	SD = 1.64	SD = 2.07
Which is more upsetting?	M = 4.18	M = 4.42
	SD = 1.65	SD = 1.92

large effect). Examining the frequencies shows that, in fact, none of the participants gave ratings of 5, 6, or 7, thus showing that none of them would feel worse about clicking the switch than dragging the photo.

However, comparing how participants felt about taking action versus not taking action, there were no significant differences in mean ratings between scenarios. The means and standard deviations are reported in Table 7. Thus, on average, participants did not think that acting *as compared to not intervening* would make them feel worse or more upset or more emotional in one scenario than in the other. (In addition, in both scenarios, the mean was close to the neutral midpoint.) However, this is consistent with the absolute level of feeling bad, upset, or emotional for *both* acting and not intervening being higher in Footbridge. It could be that participants thought that the two options in each scenario would make them feel equally bad, but the level of bad feeling was higher in Footbridge. That would be consistent with our finding that participants rated dragging the photo in Footbridge as feeling worse than clicking the switch in Side-track.

Participants also found Footbridge more emotionally arousing than Side-track. The *absolute* rating of "emotional arousal" (distance from mid-point irrespective of direction) varied between scenarios with Footbridge being more emotional than Side-track (mean of 1.79 compared to 1.27 from the neutral mid-point), $t(32) = -2.781$, $p = .009$, $d = .512$ (medium effect). This contrasts with the two questions about feeling "worse" or "upsetting", where absolute ratings did not differ. Emotional arousal goes beyond feeling bad; it can include positive emotions or even just the feeling of being "on edge". Participants may have thought that these feelings would be greater in Footbridge than in Side-track.

5. Discussion

In our real-life trolley problem, where people could choose whether or not one person should incur a small economic harm in order to save five others, by either clicking a switch (Side-track) or dragging the photo of the one (Footbridge), we found that there were no significant differences in behavior. However, the action was judged to be more right in Side-track than in Footbridge, and was also judged to be more right by observers (who did not have the possibility of taking action) than by actors (who had actually faced the choice). These differences in ratings were found in addition to differences associated with whether or not the actor actually took the action or whether the observer said that the actor should take the action. It is natural to expect that people who took action or thought it should be taken would also think that it was more right, but the positive relationship also captures any differences due to participants justifying their previous decisions. There were no significant interaction effects, which implies that the Side-track–Footbridge effect and actor-observer effect in rightness occurred both for participants who took action or said action should be taken and those who did not.

Contra the prediction of the Akrasia Hypothesis, there were no significant differences in the percentage of actors who took the action and the percentage of observers who said they should, which suggests that actors were acting in accordance with their "all-things-considered" best judgments. We did not ask actors for their all-things-considered judgments in our experiment because there would have been a clear motive to report an all-things-considered judgment that was consistent with their previous behavior. Whether there are actor-observer differences in overall judgments remains to be tested. However, in the absence of evidence for differences, we consider the all-things-considered judgments of our observers about what the actor should do to be a good proxy.

We suggest that the gap between people's moral judgments about acting and their actual behavior occurs because moral judgments about taking action are not overall judgments. Other results support this interpretation. Kurzban et al. (2012) found that in the standard version of Footbridge (their "stranger condition"), 85% of participants said that it was wrong to push the man in Footbridge but 28% of them said that they would push him; in Side-track 46% said that pulling the switch was wrong but 77% said they would pull. As with our results, there must have been participants who answered that the action was wrong but also predicted that they would act. This is a blow for any moral theory that strictly associates moral judgments with moral behavior, such as that of Hare (1952, 1963). However, Kurzban et al. also asked whether it was wrong *not* to take action, and they found that 50% of participants in Footbridge and 30% in Side-track answered "yes" to both questions, indicating that they thought it was both wrong to act and wrong not to act.

Our data might be seen as an example of a phenomenon that Williams (1981) reports in his discussion of the problem of "dirty hands". Williams argued that sometimes taking the right action may involve doing something morally discreditable, with a victim who could justly complain about having been wronged. In these cases, even though one should take action, acting still involves doing a moral wrong, which leaves a stain on the character of the actor. As applied to our data, this suggests that participants tended to think that, overall, one should take the action which leaves most children with meals, but that this still involved wronging the one child (and more so in Footbridge than in Side-track). We see this pattern of judgments amongst our observers. Hence, one might conclude, following Williams, that it is morally bad luck to find oneself in the situation of having to make a decision in Footbridge.

We found differences between actors and observers, but not the ones that are predicted by the Differing Perspectives Hypothesis. In fact, our results were the exact opposite of the predictions of the Differing Perspectives Hypothesis. There were no differences between actions and action judgments, but there was a persistent difference between the rightness judgments of our actors and observers, with observers judging the behavior as more right than actors. This is consistent with previous results found using a hypothetical trolley problem. Nadelhoffer and Feltz (2008) varied whether the decision maker was supposed to be the subject herself (actor condition) or a third person called John (observer condition), and found that 90% of the participants in the observer condition judged that it was morally permissible for John to hit the switch, but only

65% of the participants in the actor condition judged that it was morally permissible to throw the switch themselves. We replicated this result, with real agents making decisions that had real effects.

It is well known in attribution theory that there are differences between actors and observers in the explanation of behavior, with actors being more likely to cite situational constraints—the Fundamental Attribution Error (Ross & Nisbett, 1991)—so it should not come as a surprise that these show up as differences in moral judgments about self or other's actions. Consistent with the Fundamental Attribution Error, Nadelhoffer and Feltz (2008) also found that observers rated John as having more control over the outcome than actors rated themselves as having. Other research also shows that there are actor-observer differences in attributions of responsibility and freedom (Harvey, Harris, & Barnes, 1975). This suggests a possible explanation of our findings: others are regarded as having more control over the situation and hence it is more incumbent on them to act.

Another explanation for the difference in moral judgments between actors and observers draws on the dual process theory of Greene et al. (2001, 2004). According to Greene and colleagues, we have a prepotent negative emotional response to the action in Footbridge, which provokes an automatic negative moral judgment. Actors are more likely to have an emotional response than observers to the idea that they harm someone, which could cause them to make more negative moral judgments than observers.

Our results are consistent with the Separate Processing Hypothesis. Most observers judged that the actor should act and judgments that "yes" the actor should act were faster than "no" judgments. This is in contrast with evidence that we have of automatic negative *moral* judgment in Footbridge. Increasing time pressure leads to more "wrong" judgments in Footbridge-style dilemmas than in Side-track-style dilemmas (Suter & Hertwig, 2011). Cognitive load increases the time it takes to judge that acting would be morally right in Footbridge-style dilemmas, but there is no difference in time taken to make judgments that acting would be morally wrong (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). However, we found no differences in response time between Side-track and Footbridge, and no interaction effect between scenario and action judgment. Whether an interaction effect should be expected is uncertain—Greene, Sommerville, Nystrom, Darley, and Cohen (2001) claimed to find one but Greene (2009) agreed with McGuire, Langdon, Coltheart, and Mackenzie (2009) that the effect occurred only in one scenario—but the finding that "yes" action judgments are faster than "no" is the opposite of what is expected given the dual process theory of moral judgments. It is possible that moral judgment is more influenced by emotion than moral decisions, which is consistent with the hypothesis that the two processes rely on different psychological mechanisms. We do not have the response times for our participants' moral judgments, but the shape of them is consistent with the dual process theory of moral judgments, whereas neither the shape nor the timings of the action judgments are consistent with them being driven by a prepotent negative response.

We found order effects in behavior in our two tasks. Actors and observers who saw Side-track second were significantly more likely to take action than those who saw it first, leading to behavior that was more consistent with their prior response to Footbridge. Schwitzgebel and Cushman (2012) also found that responses in Side-track were more malleable than those in Footbridge. Their respondents were more likely to give the same moral goodness rating to both Footbridge and Side-track when Footbridge was presented first. We also found that significant differences in rightness ratings between Side-track and Footbridge disappeared in the second task. Our results contribute to a growing pool of evidence that judgments in trolley problems depend on the order in which they are encountered, which casts doubt on whether they can be used to provide evidence for moral principles (Liao, Wiegmann, Alexander, & Vong, 2012).

Investigations of real-life trolley problems are novel and it is hard to know what to conclude from the comparison with hypothetical life and death scenarios. Experiments with real consequences necessarily have smaller stakes and people are more risk seeking when the stakes are small (Weber & Chapman, 2005). The small stakes should tend to decrease emotional arousal but having to make a decision with real-life consequences should increase arousal. We found differences in emotional arousal for our Side-track and Footbridge scenarios but we do not know how the magnitude of the difference compares with the hypothetical life and death versions. However, despite the uncertainties involved in the comparison, real-life small stakes work is important. After all, most of our everyday moral decisions involve small stakes, so investigation of real-life small stakes decisions is important for understanding moral behavior.

Acknowledgments

The authors gratefully acknowledge support from the Arts and Humanities Research Council Grant AH/H001158/1, and from the University of Leicester for granting study leave to the second author. We thank Manisha Chauhan and Catherine Lawrence for help with data collection, Kevin McCracken for developing the software, and Walter Sinnott-Armstrong and Ming Hsu for helpful input into the experimental design. We also benefitted from discussion of the results with Robert Sugden.

References

Andreoni, J., Brown, P. M., & Vesterlund, L. (2002). What makes an allocation fair? Some experimental evidence. *Games and Economic Behavior, 40*, 1–24.
Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*, 737–753.
Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition, 10*, 381–417.
Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior, 10*, 122–142.
Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90*, 166–193.

Côté, S., Piff, P. K., & Willer, R. (2013). For whom do the ends justify the means? Social class and utilitarian moral judgment. *Journal of Personality and Social Psychology, 104*(3), 490–503.
Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics, 114*, 817–868.
Fehr, E. & Schmidt, K. (2006). The economics of fairness, reciprocity and altruism: Experimental evidence. *Handbook of the economics of giving, altruism and reciprocity* (Vol. 1). North Holland.
FeldmanHall, O., Mobbs, D., Evans, D., Hiscox, L., Navrady, L., & Dalgleish, T. (2012). What we say and what we do: The relationship between real and hypothetical moral choices. *Cognition, 123*(3), 434–441.
Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.
Frith, U., & De Vignemont, F. (2005). Egocentrism, allocentrism, and Asperger syndrome. *Consciousness and Cognition, 14*(4), 719–738.
Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making, 9*(1), 65–76.
Gold, N., Pulford, B. D., & Colman, A. M. (2013). Your money or your life: Comparing judgments in trolley problems involving economic and emotional harms, injury and death. *Economics and Philosophy, 29*, 213–233.
Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: Contextual manipulation and agent role effects in trolley problems. *Frontiers in Psychology: Cognitive Science, 5*, 35. http://dx.doi.org/10.3389/fpsyg.2014.00035.
Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology, 45*(3), 581–584.
Greene, J. D. (2014). Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics. *Ethics*.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.
Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.
Hare, R. M. (1952). *The language of morals*. Oxford: Oxford University Press.
Hare, R. M. (1963). *Freedom and reason*. Oxford: Clarendon Press.
Harvey, J. H., Harris, B., & Barnes, R. D. (1975). Actor–observer differences in the perceptions of responsibility and freedom. *Journal of Personality and Social Psychology, 32*(1), 22–28.
Hauser, M. D., Cushman, F. A., Young, L., Jin, R. K.-X., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22*(1), 1–21.
Houser, D., & Kurzban, R. (2002). Revisiting kindness and confusion in public goods experiments. *American Economic Review*, 1062–1069.
Hsu, M., Anen, C., & Quartz, S. R. (2008). The right and the good: Distributive justice and neural encoding of equity and efficiency. *Science, 320*, 1092–1095.
Koritzky, G., & Yechiam, E. (2010). On the robustness of description and experience based decision tasks to social desirability. *Journal of Behavioral Decision Making, 23*, 83–99.
Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: Pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior, 33*(4), 323–333.
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.
McGuire, J., Langdon, R., Coltheart, M., & Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *Journal of Experimental Social Psychology, 45*(3), 577–580.
Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.
Nadelhoffer, T., & Feltz, A. (2008). The actor-observer bias and moral intuitions: Adding fuel to Sinnott-Armstrong's fire. *Neuroethics, 1*(2), 133–144.
Navarrete, C. D., McDonald, M., Mott, M., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated 3-D "trolley problem". *Emotion, 12*(2), 364–370.
Osberg, T. M., & Shrauger, J. S. (1986). Self-prediction: Exploring the parameters of accuracy. *Journal of Personality and Social Psychology, 51*, 1044–1057.
Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology & Sociobiology, 17*, 145–171.
Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York: McGraw-Hill.
Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience, 18*(5), 803–817.
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.
Schwitzgebel, E., & Rust, J. (2014). The moral behavior of ethics professors: Relationships among self-reported behavior, expressed normative attitude, and directly observed behavior. *Philosophical Psychology, 27*(3), 293–327.
Sinnott-Armstrong, W. (1988). *Moral dilemmas*. Oxford and New York: Basil Blackwell.
Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition, 119*(3), 454–458.
Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology: Cognitive Science, 4*, 250. http://dx.doi.org/10.3389/fpsyg.2013.00250.
Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*, 204–217.
Thomson, J. J. (1985). The trolley problem. *Yale Law Journal, 94*, 1395–1415.
Vallone, R. P., Griffin, D. W., Lin, S., & Ross, L. (1990). Overconfident prediction of future actions and outcomes by self and others. *Journal of Personality and Social Psychology*.
Weber, B. J., & Chapman, G. B. (2005). Playing for peanuts: Why is risk seeking more common for low-stakes gambles? *Organizational Behavior and Human Decision Processes, 97*(1), 31–46.
Williams, B. (1981). *Moral luck*. Cambridge: Cambridge University Press.