11

Solving the Trolley Problem

JOSHUA D. GREENE

The Trolley Problem has baffled ethicists for decades (Foot 1978; Thomson 1985; Fischer and Ravizza 1992) and has, more recently, become a focal point for research in moral psychology (Petrinovich, O'Neill, and Jorgensen 1993; Greene *et al*. 2001; Edmonds 2013; Greene 2015). As the Trolley Problem's interdisciplinary history suggests, it is actually two closely related problems, one normative and one descriptive. The empirical research paper reprinted here (Greene *et al*. 2009) presents an approximate solution to the descriptive Trolley Problem. What's more, it may provide essential ingredients for solving – or dissolving – the normative Trolley Problem.

For the uninitiated, the Trolley Problem arises from a set of moral dilemmas, most of which involve tradeoffs between causing one death and preventing several more deaths. The descriptive problem is to explain why, as a matter of psychological fact, people tend to approve of trading one life to save several lives in some cases but not others. Consider the two most widely discussed cases (Thomson 1985): People responding to the standard *switch* case (a.k.a. *bystander*) tend to approve of hitting a switch that will redirect a trolley away from five and onto one. By contrast, people responding to the standard *footbridge* case tend to disapprove of pushing one person off a footbridge and in front a trolley, killing that person but saving five further down the track. The normative problem is to explain when and why we *ought* to approve of such one-for-many tradeoffs. The longstanding hope is that a solution to the normative Trolley Problem will reveal general moral principles. Such principles, in turn, may apply to challenging, real-world moral problems such as those encountered in the domains of bioethics (Foot 1978; Kamm 2001), war (McMahan 2009), and (most recently) the design and regulation of autonomous machines such as self-driving cars (Wallach and Allen 2008).

The normative and descriptive Trolley Problems are closely related. The normative Trolley Problem begins with the assumption that our natural responses to these cases are generally, if not uniformly, correct. Thus, any attempt to solve the normative Trolley Problem begins with an attempt to solve the descriptive problem, to identify the features of actions that elicit our moral

A Companion to Experimental Philosophy, First Edition. Edited by Justin Sytsma and Wesley Buckwalter.

approval or disapproval. Once such features have been identified and we turn toward normative questions, there are two general possibilities.

First, we might find that the features to which our judgments are sensitive also appear, upon reflection, to be features to which they *ought* to be sensitive. Under these happy circumstances, the normative problem is essentially solved. Here, we simply reconfigure our descriptive psychological principles as normative moral principles (Mikhail 2011). For example, we translate "People judge the action to be morally acceptable if and only if…" into, "The action is morally acceptable if and only if…" A philosophy thus supported would not be proven correct from first principles. Instead, it would sway comfortably in the hammock of "reflective equilibrium," supported by a network of "considered judgments" (Rawls 1971).

The second, more discomfiting possibility is that a better understanding of moral psychology will prompt us to reconsider many of our "considered judgments." More specifically, science may teach us that some of our judgments are sensitive to features that, upon reflection, do not seem to matter morally. Likewise, we may find that our judgments are insensitive to moral features that, upon reflection, do seem to matter morally. Under these more complicated circumstances, a scientific understanding of moral judgment creates a problem and a corresponding opportunity. By moving some of our judgments out of the "reliable" box and into the "unreliable" box, we may find that the ones remaining in the "reliable" box point to new conclusions. (Or to old conclusions that have been widely dismissed.)

Elsewhere I have argued that a better understanding of moral psychology favors utilitarianism/consequentialism in precisely this way (Greene 2013). My claim is not that one can derive moral "oughts" from the "is" of psychological science. Rather, the claim is that a scientific understanding of our judgments can reveal latent tensions within our preexisting set of "oughts," and thus redirect our normative thinking toward a "double-wide reflective equilibrium" (Greene 2014) – conclusions reached by incorporating scientific self-knowledge into our reflective moral theorizing. I will not defend my defense of utilitariamism/consequentialism here. Instead, my point is simply to explain how, in the most general terms, the research paper reprinted here fits into a larger project in normative ethics.

As noted earlier, the research described here provides an approximate descriptive solution to the Trolley Problem. More specifically, this research highlights the influence of two factors that exert a powerful influence when both are present. First, we are more likely to disapprove of harmful actions that involve the application of *personal force* – roughly, cases in which the agent pushes the victim. Second, we are more likely to disapprove if the harm is intended as a *means* to the agent's goal, and is not merely a foreseen (or unforeseen) side-effect.

From a normative perspective, the personal force factor is notable because it's not one that we ordinarily regard as morally relevant. Were a friend to call you from a set of trolley tracks seeking moral advice, you would probably not say, "Well, that depends. Would you have to *push* the guy, or could you do it with a switch?" The second factor, the means/side-effect factor, has a long and distinguished philosophical history (Aquinas 2006). But, as I argue elsewhere (Greene 2013), the hallowed "doctrine of double effect" may also be viewed with suspicion once its psychological origins are properly understood. Our sensitivity to the means/side-effect distinction may simply reflect the limitations of our cognitive architecture rather than a deep moral truth.

As noted earlier, the psychological theory presented in the article reprinted here is only an approximation. It's a good start, explaining much of the variability in mean ratings across the most widely discussed cases. What's more, as of this writing, I know of no theory that fits the data better. Nevertheless, several results tell us that this theory is incomplete. First, the combination of the personal force factor with the means/side-effect factor is not enough to explain the entire pattern observed in the article that follows (More specifically, it does not explain why *loop*[1] is different from *remote footbridge* and *footbridge switch*. Nor does it explain why

obstacle push is different from *standard footbridge* and *footbridge pole*.) Beyond the present data set, there are further puzzles. We know that there are (relatively weak) effects of the means/side-effect factor, even in the absence of personal force (Cushman *et al.* 2006, Schaich Borg *et al.* 2006). Likewise, we know that people react negatively to firing a fake gun at someone, even though firing a gun involves hitting a switch of sorts and nothing like pushing (Cushman *et al.* 2012). Beyond the domain of immediate bodily harm, there are nonviolent actions that seem less bad when the harm is caused indirectly and as a side effect. These include cases of damaging property (Nichols and Mallon 2006), reordering the priority list for medical treatment (Royzman and Baron 2002), and unfairly raising the price of a cancer drug (Paharia *et al.* 2009).

The most promising theory for dealing with these and other complexities is Cushman's (2013) and Crockett's (2013) account of harm-related intuition as the product of "model free" learning (Sutton and Barto 1999; Daw and Doya 2006). This theory explains how action types can acquire affective valences based on their historical consequences and how such valences can persist even when we know that the action in question will not produce the consequences that it has produced historically. Most critically for our purposes, this theory explains how our gut reactions to harmful actions can be both generally sensible and, in some cases, deeply misguided.

By confronting us with hidden truths about our minds, empirical moral psychology of the kind described in the article that follows forces moral theorists to answer tough questions: *If that's what's behind my judgment, then is my judgment worth defending? And if not, then what follows?*

Note

1 It also doesn't explain why the *collision alarm* case (Greene, 2013) differs from the *remote footbridge* and *footbridge switch* cases. This is notable because the *collision alarm* case does not involve a loop, which incorporates structural features more typical of side-effect cases. The *collision alarm* case is a more straightforward case in which harm as a means is applied in the absence of personal force. See Chapter 9 of Greene (2013) for further discussion.

References

Aquinas, T. 2006. *Summa theologiae*. Cambridge, UK: Cambridge University Press.

Edmonds, D. 2013. *Would You Kill the Fat Man?: The Trolley Problem and What Your Answer Tells Us about Right and Wrong*. Princeton, NJ: Princeton University Press.

Fischer, J. M., and M. Ravizza, eds. 1992. *Ethics: Problems and Principles*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.

Crockett, M. J. 2013. Models of morality. *Trends in Cognitive Sciences*, 17(8): 363–366.

Cushman, F. 2013. Action, Outcome, and Value A Dual-System Framework for Morality. *Personality and Social Psychology Review*, 17(3): 273–292.

Cushman, F., K. Gray, A. Gaffey, and W. B. Mendes. 2012. Simulating murder: the aversion to harmful action. *Emotion*, 12(1): 2.

Cushman, F., L. Young, and M. Hauser. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12): 1082–1089.

Daw, N. D., and K. Doya. 2006. The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2): 199–204.

Foot, P. 1978. The problem of abortion and the doctrine of double effect. In *Virtues and Vices*, 19–32. Oxford: Blackwell.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York: Penguin Press.

Greene, J. D. 2014. Beyond Point-and-Shoot Morality: Why Cognitive (Neuro) Science Matters for Ethics. *Ethics*, 124(4): 695–726.

Greene, J. D. 2015. The rise of Moral Cognition. *Cognition*, 135: 39–42.
Greene, J. D., F. A. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen. 2009. Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment. *Cognition*, 111(3): 364–371.
Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen. 2001. An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293(5537), 2105-2108.
Kamm, F. M. 2001. *Morality, Mortality: Rights, Duties, and Sstatus* (Vol. 2). Oxford, UK: Oxford University Press.
Mikhail, J. 2011. *Elements of Moral Cognition: Rawls' Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment*. Cambridge, UK: Cambridge University Press.
Nichols, S., and R. Mallon. 2006. Moral Dilemmas and Moral Rules. *Cognition*, 100(3): 530–542.
Paharia, N., K. S. Kassam, J. D. Greene, and M. H. Bazerman. 2009. Dirty Work, Clean Hands: The Moral Psychology of Indirect Agency. *Organizational Behavior and Human Decision Processes*, 109(2): 134–141.
Petrinovich, L., P. O'Neill, and M. Jorgensen. 1993. *An Empirical Study of*
Moral Intuitions: Toward an Evolutionary Ethics. *Journal of Personality and Social Psychology* 64: 467–478.
Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Royzman, E. B., and J. Baron. 2002. The Preference for Indirect Harm. *Social Justice Research*, 15: 165–184.
Schaich Borg, J., C. Hynes, J. Van Horn, S. Grafton, and W. Sinnott-Armstrong. 2006. Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, 18(5): 803–817.
Sutton, R. S., and A. G. Barto. 1999. Reinforcement Learning. *Journal of Cognitive Neuroscience*, 11(1): 126–134.
Thomson, J. 1985. The Trolley Problem. *Yale Law Journal*, 94(6): 1395–1415.
Wallach, W., and C. Allen. 2008. *Moral Machines: Teaching Robots Right from Wrong*. Oxford, UK: Oxford University Press.

*Pushing Moral Buttons: The Interaction between Personal Force and Intention in Moral Judgment**

JOSHUA D. GREENE, FIERY A. CUSHMAN, LEIGH E. NYSTROM, LISA E. STEWART, KELLY LOWENBERG, AND JONAT HAN D. COHEN

1 Introduction

Many moral and political controversies involve a tension between individual rights and the greater good (Singer, 1979). This tension is nicely captured by a puzzle known as the "trolley problem" that has long interested philosophers (Foot, 1978; Thomson, 1985) and that has recently become a topic of sustained neuroscientific (Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005; Schaich Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006) and psychological (Cushman, Young, & Hauser, 2006; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000, 2007; Moore, Clark, & Kane, 2008; Nichols & Mallon, 2005; Waldmann & Dieterich, 2007) investigation. One version of the trolley problem is as follows: A runaway trolley is about to run over and kill five people. In the switch dilemma[1] one can save them by hitting a switch that will divert the trolley onto a side-track, where it will kill only one person. In the footbridge dilemma one can save them by pushing someone off a footbridge and into the trolley's path, killing him, but stopping the trolley. Most people approve of the five-for-one tradeoff in the switch dilemma, but not in the footbridge dilemma (Cushman, Young, & Hauser, 2006; Greene et al., 2001; Petrinovich, O'Neill, & Jorgensen, 1993).

What explains this pattern of judgment? Neuroimaging (Greene et al., 2001, 2004), lesion (Ciaramelli et al., 2007; Koenigs et al., 2007; Mendez et al., 2005), and behavioral (Bartels, 2008; Greene et al., 2008; Valdesolo & DeSteno, 2006) studies indicate that people respond differently to these two cases because the action in the footbridge dilemma elicits a stronger negative emotional response. But what features of this action elicit this response? Recent studies implicate two general factors. First, following Aquinas (2006), many appeal to intention and, more specifically, the distinction between harm intended as a means to a greater good (as in the footbridge dilemma) and harm that is a foreseen but "unintended" side-effect of achieving a greater good (as in the switch dilemma) (Cushman et al., 2006; Hauser, Cushman, Young, Jin, & Mikhail, 2007; Mikhail, 2000; Schaich Borg et al., 2006). Second, many studies appeal to varying forms of "directness" or

* The article was originally published in Cognition in 2009, reprinted with permission

"personalness," including physical contact between agent and victim (Cushman et al., 2006), the locus of intervention (victim vs. threat) in the action's underlying causal model (Waldmann & Dieterich, 2007), whether the action involves deflecting an existing threat (Greene et al., 2001), and whether the harmful action is mechanically mediated (Moore et al., 2008; Royzman & Baron, 2002). The aim of this paper is to integrate these two lines of research.

We present two experiments examining a directness/personalness factor that we call personal force. An agent applies personal force to another when the force that directly impacts the other is generated by the agent's muscles, as when one pushes another with one's hands or with a rigid object. Thus, applications of personal force, so defined, cannot be mediated by mechanisms that respond to the agent's muscular force by releasing or generating a different kind of force and applying it to the other person. Although all voluntary actions that affect others involve muscular contractions, they do not necessarily involve the application of personal force to another person. For example, firing a gun at someone or dropping a weight onto someone by releasing a lever do not involve the application of personal force because the victims in such cases are directly impacted by a force that is distinct from the agent's muscular force, i.e. by the force of an explosion or gravity. The cases of direct harm examined by Royzman and Baron (2002) are not so direct as to involve the application of personal force. The direct/indirect distinction described by Moore and colleagues (2008) is similar to the distinction drawn here between personal and impersonal force, but Moore and colleagues do not systematically distinguish between physical contact and personal force.

Experiments 1a and b aim to document the influence of personal force, contrasting its effect with those of physical contact (1a–b) and spatial proximity (1a) between agent and victim. Experiment 1a also introduces a method for controlling for effects of unconscious realism, i.e. a tendency to unconsciously replace a moral dilemma's unrealistic assumptions with more realistic ones. ("Trying to stop a trolley with a person is unlikely to work.") Experiments 2a and b examine the interaction between personal force and intention. More specifically, we ask whether the effect of personal force depends on intention and vice versa.

2 Experiment 1a

We compared four versions of the footbridge dilemma to isolate the effects of spatial proximity, physical contact, and personal force on moral judgments concerning harmful actions. We also tested the unconscious realism hypothesis by controlling for subjects' real-world expectations.

2.1 Method

2.1.1 Subjects

Subjects were 271 females, 337 males, and 12 genders unknown. The mean age was 31. Subjects were recruited anonymously in public venues in New York City and Boston. Subjects were paid $3.

2.1.2 Design, materials, and procedure

Subjects responded to one of four versions of the footbridge dilemma in a between-subject design, indicating the extent to which the proposed action is "morally acceptable." In the standard footbridge dilemma (n = 154, Fig. 1a), the agent (named Joe) may save the five by pushing the victim off the footbridge using his hands. This action involves spatial proximity, physical contact, and personal force. In the remote footbridge dilemma (n = 82, Fig. 1d), Joe may drop the victim onto the tracks using a trap door and a remote switch. This action involves none of the three aforementioned factors. The footbridge pole dilemma (n = 72, Fig. 1b) is identical to the standard footbridge

dilemma except that Joe uses a pole rather than his hands to push the victim. This dilemma involves spatial proximity and personal force without physical contact. The footbridge switch dilemma (n = 160, Fig. 1c) is identical to the remote footbridge dilemma except that Joe and the switch are adjacent to the victim. This dilemma involves spatial proximity without physical contact or personal force. Comparing remote footbridge to footbridge switch isolates the effect of spatial proximity. Comparing standard footbridge to footbridge pole isolates the effect of physical contact. Comparing footbridge switch to footbridge pole isolates the effect of personal force.

The text of the standard footbridge dilemma is as follows:

An empty runaway trolley is speeding down a set of tracks toward five railway workmen. There is a footbridge above the tracks in between the runaway trolley and the five workmen. On this footbridge is a railway workman wearing a large, heavy backpack. If nothing is done, the trolley will proceed down the main tracks and cause the deaths of the five workmen (see Fig. 1).

It is possible to avoid these five deaths. Joe is a bystander who understands what is going on and who happens to be standing right behind the workman on the footbridge. Joe sees that he can avoid the deaths of the five workmen by pushing the workman with the heavy backpack off of the footbridge and onto the tracks below. The trolley will collide with the workman, and the combined weight of the workman and the backpack will be enough to stop the trolley, avoiding the deaths of the five workmen. But the collision will cause the death of the workman with the backpack.

Is it morally acceptable for Joe to push the workman off of the footbridge in order to avoid the deaths of the five workmen, causing the death of the single workman instead?

Subjects answered (YES/NO) and rated the moral acceptability of the action on a nine-point scale. The above text was accompanied by a diagram (Fig. 1a). Similar text and diagrams (Figs. 1c–d and Fig. 3) were used for other dilemmas, with changes reflecting the experimental manipulations. Complete materials are available at (url: https://mcl.wjh.harvard.edu/materials/Greene-Cogn09-SuppMats.pdf).

The instructions acknowledged that the dilemmas were not necessarily realistic and requested that subjects "suspend disbelief." Data from 31 (of 664) subjects who reported being unable/

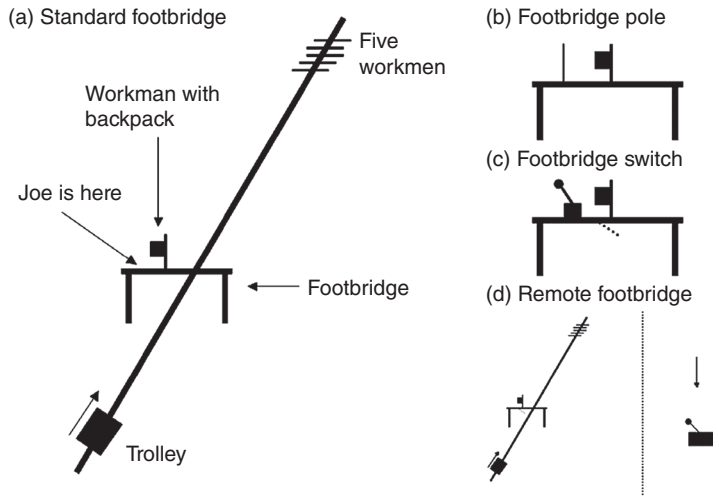

Figure 1 Diagrams for the (a) standard footbridge dilemma (physical contact, spatial proximity, and personal force); (b) footbridge pole dilemma (spatial proximity and personal force); (c) footbridge switch dilemma (spatial proximity); and (d) remote footbridge dilemma. (Panels b–d depicts details of diagrams presented to subjects with labels and some pictorial elements removed for clarity.)

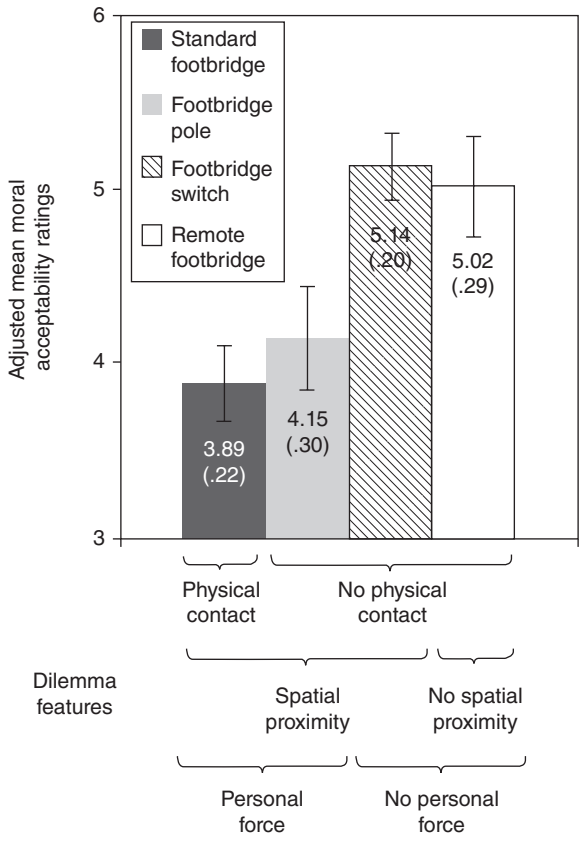

Figure 2 Results of Experiments 1a and b: Moral acceptability ratings for four dilemmas in which the proposed harmful actions vary in their involvement of physical contact, spatial proximity, and personal force. Error bars indicate SEM. Numbers within graph bars indicate mean and SEM, adjusted for effects of covariates. Note: Joe cannot avoid the deaths of the five workmen by jumping himself because he is not heavy enough to stop the trolley. There is also not enough time to remove the backpack from the workman.

unwilling to suspend disbelief ("conscious realists") were excluded form analysis, as were data from 10 subjects reporting confusion.

To control for unconscious realism, we asked subjects (after they responded to the dilemma) to report on their real-world expectations concerning the likely consequences of Joe's actions. Subjects estimated the likelihood (0–100%) that the consequences of Joe's action would be (a) as described in the dilemma (five lives saved at the cost of one), (b) worse than this, or (c) better than this. These estimates (respectively, labeled PLAN, WORSE, and BETTER) were modeled as covariates. The predictive value of these variables indicates the extent to which subjects' judgments may reflect unconscious realism.

Data were analyzed using a general linear model. Here and in Experiment 2a, the three "realism covariates" and gender were included as first-order covariates and allowed to interact with the dilemma variable. In Experiment 2a these factors were allowed to interact with both main effects and the interaction of interest. Because the realism covariates are likely correlated, this analysis is adequate to control for their collective effects but inadequate to resolve their respective contributions.

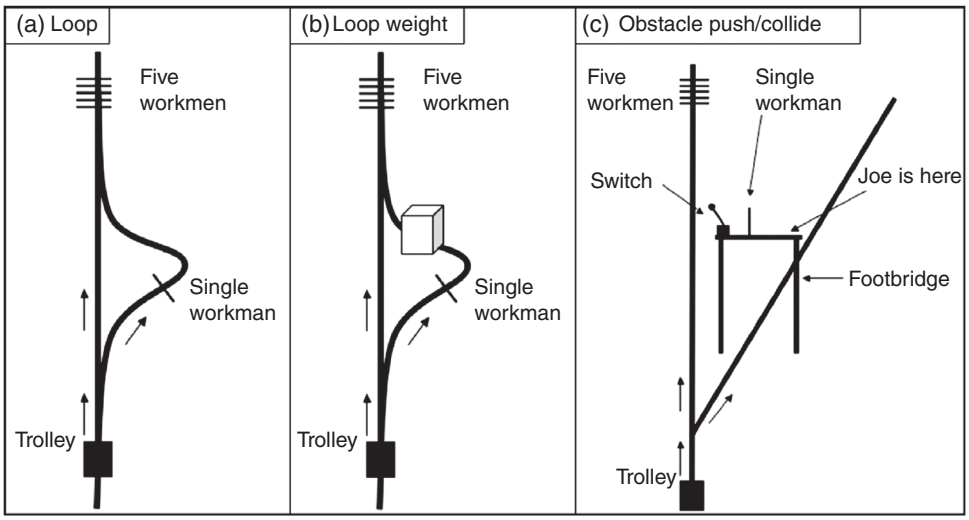

Figure 3 Diagrams for the (a) loop dilemma (means, no personal force); (b) loop weight dilemma (side-effect, no personal force); (c) obstacle push dilemma (means, personal force), and obstacle collide dilemma (side-effect, personal force). Remote switches (as in Fig. 1d) not showed in panels a–b.

2.2 Results

Ratings of the moral acceptability of sacrificing one life to save five differed among the four dilemmas ($F(3, 417) = 9.69$, $p < 0.0001$). Planned pairwise contrasts revealed no significant effect of spatial proximity (remote footbridge vs. footbridge switch: $F(1, 417) = 0.11$ $p = 0.74$), no significant effect of physical contact (standard footbridge vs. footbridge pole: $F(1, 417) = 1.43$. $p = 0.23$), but a significant effect of personal force (footbridge switch vs. footbridge pole: $F(1, 417) = 7.63$, $p = 0.006$, $d = 0.40$; see Fig. 2). There was a significant main effect of WORSE ($F(1, 417) = 5.80$, $p = 0.02$) with actions expected to be less successful eliciting lower moral acceptability ratings, consistent with unconscious realism. There were no significant effects of PLAN, BETTER, gender, or higher order covariates ($p > 0.05$).

These results indicate that harmful actions involving personal force are judged to be less morally acceptable. Moreover, they suggest that spatial proximity and physical contact between agent and victim have no effect and that a previously reported effect of physical contact (Cushman et al., 2006) is in fact an effect of personal force. In all four of the dilemmas examined in this study, the harmful event is intended as a means to achieving the agent's goal, raising the possibility that the effect of personal force is limited to cases in which the harm is intended as a means. Experiments 2a and b examine the interaction between personal force and intention.

3 Experiment 1b

To ensure that the results concerning personal force and physical contact observed in Experiment 1a generalize to other contexts, we conducted an additional experiment using a different set of moral dilemmas, as well as a different rating scale.

3.1 Method

3.1.1 Subjects

Subjects were 54 females and 37 males, with a mean age of 31. Subjects were unpaid and recruited anonymously through the Alkami Biobehavioral Institute's Research Subject Volunteer Program (http://rsvp.alkami.org/), psychological research on the net (http://psych.han-over.edu/Research/exponnet.html), and craigslist (http://www.craigslist.org). Subjects participated through the Greene/Moral Cognition Lab's online research page: https://mcl.wjh.harvard.edu/online.html.

3.1.2 Design, materials, and procedure

Subjects responded to one of three versions of the speedboat dilemma (Cushman et al., 2006), in which saving the lives of five drowning swimmers requires lightening the load of a speedboat. This requires removing from the speedboat a passenger who cannot swim, causing that passenger to drown. In the first version (Pc–Pf), the agent pushes the victim with his hands, employing physical contact and personal force. In the second version (NoPc–Pf), the agent pushes the victim with an oar, employing personal force, but no physical contact. In the third version (NoPc–NoPf), the agent removes the victim by accelerating quickly, causing the victim to tumble off the back of the boat. This employs neither personal force nor body contact. Following Cushman et al. (2006), subjects evaluated the agent's action using a seven-point scale with one labeled "Forbidden," four labeled "Permissible," and seven labeled "Obligatory".

3.2 Results

Ratings varied significantly among the three dilemmas (M (SD) for Pc–Pf = 2.28 (1.50); NoPc–Pf = 2.33 (1.20); NoPc–NoPf = 3.3 (1.58); $F(2, 87) = 4.72$, $p = 0.01$). As predicted, planned contrasts revealed no significant effect of physical contact (Pc–Pf vs. NoPc–Pf: $F(1, 87) = 0.02$ $p = 0.89$), but a significant effect of personal force (NoPc–Pf vs. NoPc–NoPf: $F(1, 87) = 5.86$, $p = 0.02$, $d = 0.69$).

4 Experiment 2a

This experiment examined the independent effects of personal force and intention and, most critically, their interaction, by comparing four dilemmas using a 2 (personal force absent vs. present) × 2 (means vs. side-effect) design.

4.1 Method

Methods follow Experiment 1a unless otherwise noted.

4.1.1 Subjects

Subjects were 181 females, 179 males, and 6 genders unknown. Mean age: 31. An additional 44 subjects were excluded for "realism"/confusion.

4.1.2 Design, materials, and procedure

Each subject responded to one of four dilemmas. In the loop dilemma (Hauser et al., 2007; Mikhail, 2000; Thomson, 1985; Waldmann & Dieterich, 2007), Joe may save the five by turning the trolley onto a looped side-track that reconnects with the main track at a point before the five people ($n = 152$, Fig. 3a). There is a single person on the side-track who will be killed if the trolley is turned, but who will prevent the trolley from looping back and killing the five. Here the victim is harmed as

a means (i.e. intentionally), but without the application of personal force. The loop weight dilemma (Hauser et al., 2007; Mikhail, 2000) is identical to the loop dilemma except that a heavy weight positioned behind the victim on the side-track, rather than the victim, stops the trolley (n = 74, Fig. 3b). Here the victim is killed as a side-effect (i.e. without intention) and, again, without the application of personal force. In the obstacle collide dilemma, the victim is positioned on a high and narrow footbridge in between Joe and a switch that must be hit in order to turn the trolley and save the five (n = 70, Fig. 3c). To reach the switch in time, Joe must run across the footbridge, which will, as a side-effect, involve his colliding with the victim, knocking him off the footbridge and to his death. Thus, this dilemma involves personal force, but not intention. The obstacle push dilemma (n = 70) is identical to the obstacle collide dilemma except that Joe must push the victim out of the way in order to get to the switch. Although the victim is not used to stop the trolley, Joe performs a distinct body movement (pushing) that is both harmful and necessary for the achievement of the goal. Thus, this dilemma involves the application of personal force that is intentional.

4.2 Results

There was a main effect of intention (loop and obstacle push vs. loop weight and obstacle collide: $F(1, 329) = 6.47$, $p = 0.01$) and no main effect of personal force (loop dilemmas vs. obstacle dilemmas: $F(1, 329) = 4.85$, $p = 0.29$). Crucially, we observed the predicted interaction between intention and personal force ($F(1, 329) = 7.54$, $p = 0.006$, partial $\eta^2 = 0.02$). A series of planned pairwise contrasts clarified the nature of this interaction: Comparing the loop, loop weight, and obstacle collide dilemmas revealed no significant effects ($p > 0.2$), while the obstacle push dilemma elicited significantly lower moral acceptability ratings than each of these other dilemmas (obstacle push vs. others, respectively: $F(1, 329) = 8.20$, 5.56, and 11.85; $p = 0.004$, 0.02, 0.0006) (see Fig. 4). This suggests that the main effect of intention reported above is explained by the conjoint effect of personal force and intention (i.e. by the uniquely low moral acceptability ratings elicited by the obstacle push dilemma). There were significant effects of WORSE ($F(1, 329) = 15.80$, $p < 0.0001$) and PLAN ($F(1, 329) = 19.21$, $p < 0.0001$). Males tended toward higher moral acceptability ratings ($F(1, 329) = 4.99$, $p = 0.03$), particularly in the absence of personal force (gender × personal force: ($F(1, 329) = 6.54$, $p = 0.01$). There was no significant effect of BETTER or other higher order covariates ($p > 0.05$).

5 Experiment 2b

To ensure that the main results observed in Experiment 2a generalize to other contexts, we recoded and reanalyzed the data from Cushman et al. (2006). More specifically, we examined the moral permissibility ratings for the 19 moral dilemmas involving actions (rather than omissions), including five dilemmas in which the harm is caused as a means without personal force (Means–noPf), six dilemmas in which the harm is caused as a side-effect without personal force (SE–noPf), three dilemmas in which the harm is caused as a means with personal force (Means–Pf), and five dilemmas in which the harm is caused as a side-effect with personal force (SE–Pf). Dilemma codings followed those of Cushman et al., with personal force replacing physical contact, except that two dilemmas not involving physical contact were deemed (prior to analysis) to involve personal force. (See online Supplementary materials). Because our interest here is in testing the generalizability of our results across contexts, we used dilemma/item, rather than subject, as the unit of analysis.

Ratings varied significantly among the four dilemma types (M (SD) for Means–noPf = 3.58 (0.55); SE–noPf = 4.25 (0.37); Means–Pf = 2.92 (0.44); SE–Pf = 4.53 (0.35); $F(3, 15) = 10.93$, $p = 0.0005$). There was a main effect of intention: $F(1, 15) = 31.08$, $p < 0.0001$) and

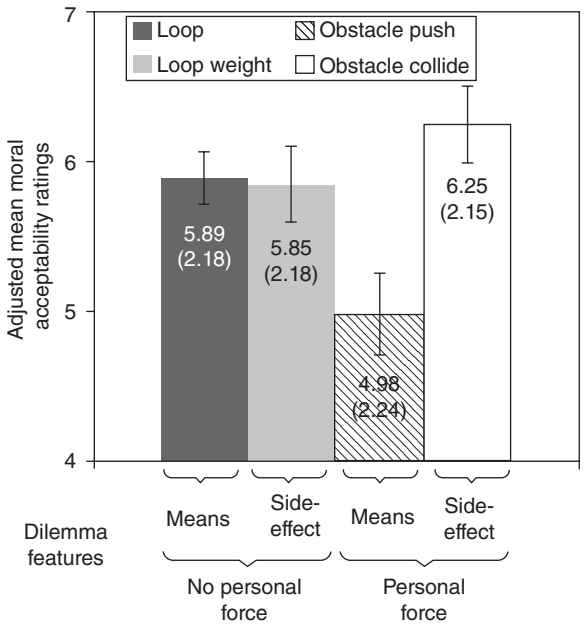

Figure 4 Results of Experiments 2a and b: Moral acceptability ratings for four dilemmas in which the proposed harmful actions vary in their intentional status (means vs. side-effect) and the presence/absence of personal force. Error bars indicate SEM. Numbers within graph bars indicate mean and standard deviation, adjusted for effects of covariates.

no main effect of personal force ($F(1, 15) = 0.90$, $p = 0.36$). Crucially, we observed the predicted interaction between intention and personal force ($F(1, 15) = 5.35$, $p = 0.04$, partial $\eta^2 = 0.26$). As predicted, the simple effect of personal force was significant when the harm was a means ($F(1, 15) = 4.49$, $p = 0.05$), but not when the harm was a side-effect ($F(1, 15) = 1.14$, $p = 0.30$), indicating that the effect of personal force depends on intention. In this experiment, however, the effect of intention was not only significant in the presence of personal force ($F(1, 15) = 26.24$, $p = 0.0001$), but also in the absence of personal force, albeit more weakly ($F(1, 15) = 6.43$, $p = 0.02$).

6 Discussion

In two sets of experiments, harmful actions were judged to be less morally acceptable when the agent applied personal force to the victim. In Experiments 1a and b the effect of personal force was documented and distinguished from effects of physical contact (Cushman et al., 2006) and spatial proximity (1a only), which were not significant. Experiments 2a and b revealed that personal force interacts with intention, such that the personal force factor only affects moral judgments of intended harms, while the intention factor is enhanced in cases involving personal force. Put simply, something special happens when intention and personal force co-occur. (We note that all key results held using categorical (YES/NO) judgments when they were collected.)

In Experiments 2a and b, personal force exhibited no effect in the absence of intention, a striking result in light of Experiments 1a and b and previous work. In Experiment 2a, the action in the obstacle collide dilemma was judged to be as acceptable as those in the loop, and loop

weight dilemmas despite the fact that obstacle collide, unlike the other two dilemmas, involves direct harm (Moore et al., 2008; Royzman & Baron, 2002), physical contact (Cushman et al., 2006), harm not caused by the deflection of an existing threat (Greene et al., 2001), and an alteration of the victim's causal path (Waldmann & Dieterich, 2007). (One may interpret Waldmann & Dieterich as assuming that victim interventions are necessarily intended, in which case this result is consistent with their theory.) Experiment 2b showed that this finding generalizes to several additional dilemma contexts, strongly suggesting that the effect of personal force is limited to cases involving harm as a means.

Experiments 2a and b also demonstrate that the effect of the intention factor on moral judgment is enhanced in cases involving personal force, and Experiment 2a found no effect of intention in the absence of personal force, suggesting that intention operates only in conjunction with other factors such as, but not necessarily limited to, personal force. Our finding of equivalence between the loop (intentional harm) and loop weight (harmful side-effect) dilemmas directly contradicts some earlier findings (Hauser et al., 2007; Mikhail, 2000),[2] but is consistent with other earlier findings (Waldmann & Dieterich, 2007). Following Waldmann & Dieterich, we attribute the effects observed by Hauser et al. (2007) and Mikhail (2000) to a confound whereby the loop dilemma, but not the loop weight dilemma, refers to the victim as a "heavy object." ("There is a heavy object on the side-track… The heavy object is 1 man…" vs. "There is a heavy object on the side-track… There is 1 man standing on the side-track in front of the heavy object…").

The statistical significance of the "unconscious realism" covariates included in Experiments 1a and 2a provides limited support for the unconscious realism hypothesis. This support is limited for at least two reasons. First, subjects' assessments of the likely real-world effects of the actions in question may be post–hoc rationalizations (Haidt, 2001). Second, a correlation between real-world expectations and moral judgments is not sufficient to establish a causal relationship. Nevertheless, these results indicate that effects of unconscious realism may be real and that researchers who use hypothetical cases to study decision-making should consider controlling for such effects as done here.

One might wonder why the actions judged to be more acceptable in Experiment 1a (footbridge switch and remote footbridge) received comparable ratings (~5) to the action judged to be less acceptable in Experiment 2a (obstacle push). First, in considering why the footbridge switch and remote footbridge dilemmas received relatively low ratings, we speculate that this may be due to the fact that the actions in these dilemmas involve dropping the victim onto the tracks, constituting an additional intentional harm (Mikhail, 2007). Second, in considering why the ratings for the obstacle push dilemma are relatively high, we suggest that this may be due to the fact that the action in the obstacle push dilemma, while involving a distinct body movement that is harmful and necessary for the achievement of the goal, does not involve using the victim, as in the four footbridge dilemmas. Each of these hypotheses will be explored in future work.

The latter hypothesis highlights more general open questions concerning the scope of agents' intentions (Bennett, 1995). In the obstacle push dilemma, the pushing is necessary, but the consequent harm, strictly speaking, is not. This observation raises parallel questions about more paradigmatic cases of intentional harm. For example, one might claim that even in the standard footbridge dilemma the harm is unintentional because the agent merely intends to use the victim's body to stop the trolley, harming him only as a foreseen side-effect of doing this. These observations highlight the need for a theory of intentional event segmentation (Zacks & Tversky, 2001).

Other open questions concern the proper characterization of personal force: Must it be continuous (as in pushing), or may it be ballistic (as in throwing)? Is pulling equivalent to pushing? We acknowledge, more broadly, that the effects documented here under the rubric of "personal force" may ultimately be refined and reinterpreted. For example, alternative interpretations may focus on the potential for dynamic interaction between agent and victim.

Finally, we consider the significance of our finding that personal force and intention interact: Why is it that the combined presence of personal force and intention pushes our moral buttons? The co-dependence of these factors suggests a system of moral judgment that operates over an integrated representation of goals and personal force—representations such as "goal-within-the-reach-of-muscle-force." In a general sense, this suggests a mechanism of moral judgment that is a species of embodied cognition (Gallese, Keysers, & Rizzolatti, 2004; Lakoff & Johnson, 1999; Prinz, 2002; Wilson, 2002). One natural source of such embodied goal representations is system of action planning that coordinates the application of personal force to objects to achieve goal-states for those specific objects. A putative sub-system of moral judgment, monitoring such action plans, might operate by rejecting any plan that entails harm as a goal-state (Mikhail, 2000; Mikhail, 2007) to be achieved through the direct application of personal force. We propose this "action-planning" account of the present results as an important area for further research.

At a more general level, the present study strongly suggests that our sense of an action's moral wrongness is tethered to its more basic motor properties, and specifically that the intention factor is intimately bound up with our sensitivity to personal force. This perspective contrasts with at least some versions of the "universal moral grammar" perspective (Hauser, 2006; Mikhail, 2000; Mikhail, 2007), according to which the present moral judgments depend on goal representations of the kind one might find in a legal system, leaving little room for an 'embodied' representation involving personal force. It also presents a challenge to philosophical theories that endorse the doctrine of double effect (i.e. the intention factor) on the basis of its intuitive plausibility (Aquinas, 2006; Fischer & Ravizza, 1992). Will they bless its shotgun marriage to a normatively ugly bride: the doctrine of personal force?

Acknowledgements

We thank Andrew Conway, Daniel Gilbert, Andrea Heberlein, Wendy Mendes, and Daniel Wegner for their assistance. This work was supported by the NSF (BCS-0351996) and NIH (MH067410).

Appendix: Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.cognition. DOI:10.1016/j.cognition.2009.02.001 2009.02.001 DOI:10.1016/j.cognition.2009.02.001 .

Notes

1 Previously we have referred to this as the "trolley" dilemma (Greene et al., 2001).

2 This analysis had adequate power (0.97) to detect a small effect ($d = 0.2$) trending weakly ($p < 0.95$) in the predicted direction, but none was observed.

References

Aquinas, T. (2006). *Summa theologiae*. Cambridge University Press.

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

Bennett, J. (1995). *The act itself*. New York: Oxford University Press.
Fischer, J. M., & Ravizza, M. (Eds.). (1992). *Ethics: Problems and principles*. Fort Worth, TX: Harcourt Brace Jovanovich College Publishers.
Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, 2(2), 84–92.
Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, 17(12), 1082–1089.
Foot, P. (1978). *The problem of abortion and the doctrine of double effect. In Virtues and vices*. Oxford: Blackwell.
Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8(9), 396–403.
Greene, J., Morelli, S., Lowenberg, K., Nystrom, L., & Cohen, J. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, 107(3), 1144–1154.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, 44(2), 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105–2108.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108, 814–834.
Hauser, M. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco.
Hauser, M., Cushman, F., Young, L., Jin, R. K., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind and Language*, 22(1), 1–21.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, 446(7138), 908–911.
Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.
Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, 18(4), 193–197.
Mikhail, J. (2000). *Rawls' linguistic analogy: A study of the "generative grammar" model of moral theory described by John Rawls in a theory of justice, unpublished doctoral dissertation*. Cornell University
Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, 11(4), 143–152.
Moore, A., Clark, B., & Kane, M. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, 19(6), 549–557.
Nichols, S., & Mallon, R. (2005). *Moral dilemmas and moral rules*. Cognition.
Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward an evolutionary ethics. *Journal of Personality and Social Psychology*, 64, 467–478.
Prinz, J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. Cambridge, MA: MIT.
Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, 15, 165–184.
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.
Singer, P. (1979). *Practical ethics*. Cambridge: Cambridge University Press.
Thomson, J. (1985). The trolley problem. *Yale Law Journal*, 94(6), 1395–1415.
Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, 17(6), 476–477.
Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, 18(3), 247–253.
Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636.
Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127(1), 3–21.