A single counterexample leads to moral belief revision

Zachary Horne
Departments of Psychology and Philosophy, University of Illinois at Urbana-Champaign

Derek Powell
Department of Psychology, University of California, Los Angeles

John Hummel
Department of Psychology, University of Illinois at Urbana-Champaign

Running head: A single counterexample leads to moral belief revision

Address for correspondence:

Zachary Horne
Department of Psychology
University of Illinois, Urbana-Champaign
603 E. Daniel St.
Champaign, IL 61820
Email: horne2@illinois.edu

Author Note

A preliminary report of Experiment 1 was presented at the 35th Annual Conference of the Cognitive Science Society (Berlin, August 2013). Correspondence may be addressed to Zachary Horne, Department of Psychology, University of Illinois, Urbana-Champaign, 603 E. Daniel St. Champaign, IL 61820, email: horne2@illinois.edu

Keywords: *belief revision; morality; moral conviction; confirmation bias; moral dilemmas*

Abstract

What kind of evidence will lead people to revise their moral beliefs? Moral beliefs are often strongly held convictions, and existing research has shown that morality is rooted in emotion and socialization rather than deliberative reasoning. Additionally, more general issues – such as confirmation bias – further impede coherent belief revision. Here we explored a unique means

for inducing belief revision. In two experiments, participants considered a moral dilemma in which an overwhelming majority of people judged that it was inappropriate to take action to maximize utility. Their judgments contradicted a utilitarian principle they otherwise strongly endorsed. Exposure to this scenario led participants to revise their belief in the utilitarian principle, and this revision persisted over several hours. This method provides a new avenue for inducing belief revision.

A single counterexample leads to moral belief revision

Introduction

Conventional wisdom advises one to steer clear of religion and politics in polite conversation. One reason for avoiding these topics is that they can quickly turn to disagreements about what is moral. Such disagreements are often fruitless: people's moral beliefs are highly resistant to revision (Skitka, 2010) as they are thought to be rooted in emotion, motivation, and socialization (e.g., Greene, 2007; Haidt, 2001; Prinz, 2007; Rai & Fiske, 2012). Some of these roots go deep: there is evidence that infants have moral beliefs as early as 19 months of age (e.g., Sloane, Baillargeon, & Premack, 2012; also see Hamlin, Wynn, & Bloom, 2007), and have developing moral frameworks by age five (e.g., Kohlberg, 1976; Rochat et al., 2009; Turiel, 2006). Moral beliefs may be particularly steadfast because, as Skitka (2010) observes, "to support alternatives to what is right, moral and good is to be absolutely wrong, immoral, if not evil". Yet, people can and often do change their moral beliefs, so it seems that research on moral conviction and moral development bring a puzzling psychological picture into focus. Moral beliefs are at once stubbornly intransigent and resistant to revision, and yet their change and development throughout people's lifespan is almost inevitable.

Despite the inevitability of moral development, assimilating new evidence into one's moral framework is not a trivial task. The vast literature on confirmation bias has shown that

people have an overwhelming tendency to seek out evidence that confirms their beliefs, and to interpret any new evidence (even evidence objectively contrary to their beliefs) as confirmatory (e.g., Klayman, 1995; Nickerson, 1998). This resistance to change has also been observed in the moral domain: for instance, people do not temper their credences about the death penalty in the face of compelling contradictory evidence. Rather, they often view evidence that contradicts their belief as confirmatory, leading to even more polarized attitudes (e.g., Lord, Ross, & Lepper, 1979).

There are, of course, instances where people do revise even their deeply held beliefs. For a particularly notable example, consider former United States Federal Reserve chair Alan Greenspan, who changed many of his views on U.S. government regulation in light of the economic crises of 2008 (*Time*, October 23, 2008). The overarching point of confirmation bias is not that people never change their minds -- just that it takes more evidential support than seems (optimally) warranted, such as a global economic collapse. Even here, Greenspan altered his beliefs about economic policy, but there is reason to think his attitudes about the *moral* or normative aspects of economic policy did not change. In describing his new position on increasing market regulation, Greenspan prefaced by saying, "*As much as I would prefer it otherwise*, in this financial environment I see no choice …" (emphasis ours). All of this research raises the question, what kind of evidence can cause people to revise their moral beliefs?

We propose that the method of cases, a common practice in philosophy (Nagel, 2012), provides an avenue for approaching this question. In ethical theory, moral dilemmas (a kind of case) are constructed with the aim of advancing or countering moral theories. Consider the Transplant dilemma, a case in which one must decide whether to harvest a single person's organs in order to save the lives of five dying patients. Many philosophers think it is intuitively wrong to

kill the single person, even to save the many (e.g., Kamm, 2007; Thomson, 1976), and therefore see their judgments about this dilemma as inconsistent with utilitarianism.[1] Likewise, we hypothesized that people would view their judgments about the Transplant dilemma as being inconsistent with some of their moral beliefs (for instance, the belief that you should always save the most lives). If so, then exposure to this dilemma could lead to moral belief revision.

Coherence, the method of cases, and belief revision

An important epistemic standard for evaluating a belief is how it coheres with one's other beliefs. We think that coherence may be particularly important for ethics, where there is little hope of independent and objective verification of beliefs through, say, perception. Is there any evidence that people value coherence when evaluating an epistemic agent? First, extant research suggests that striving for coherence often leads to cognitive dissonance, a downstream consequence of which is often belief revision (see Cooper, 2007 for a review). Second, people's distaste for hypocrisy also suggests a certain level of concern for coherence. For example, consider the U.S. congressman from Florida who was recently charged with cocaine possession. The public seemed less concerned with his crime (a misdemeanor) than with his prior vote in favor of a drug-testing requirement for welfare recipients (New York Times, January 27, 2014). More than his drug use, it was his inconsistency that was repugnant.

Still, maintaining coherence in one's system of moral beliefs is not always easy. People hold a multitude of moral beliefs, some portion of which are recruited when making judgments in any particular moral situation (e.g., Nichols & Mallon, 2006; Nichols, Kumar, & Lopez, Draft). It's possible that incompatibilities between stored moral beliefs could go largely unnoticed, as typically only a handful of beliefs may ever be actively represented together. In

[1] Crudely put, the view that one should maximize utility. However, see Parfit (2011) for an argument that philosopher's intuitions about this dilemma do not actually run contrary to utilitarianism.

contrast, considering moral dilemmas pits opposing attitudes against one another, making salient the tensions between the beliefs one recruits when considering these dilemmas (Nichols, 2004; Prinz, 2007). Therefore, we predict that making judgments about moral dilemmas will lead to moral belief revision by encouraging people to be coherent.

This account suggests that considering a moral dilemma is quite different from considering the sorts of evidence typically used in research on confirmation bias. In these studies, participants are often presented with external evidence (e.g., a relevant statistic about the efficacy of the death penalty at deterring murder) that contradicts a belief they hold. Faced with "external conflict", participants tend to either reject the presented evidence, or distort it until it conforms to their beliefs. However, when considering a case like the Transplant dilemma, the tension results due to a conflict between the participants' *own* beliefs. This difference, between internal and external conflict, motivates our prediction that considering a moral dilemma will lead to moral belief revision, even though "external" evidence typically does not.

Consistent with this prediction, one recent study suggests that people revise their beliefs about the impermissibility of killing after considering moral dilemmas. Horne, Powell, and Spino (2013) asked participants to make a judgment about a moral dilemma and then immediately after rate their agreement with different moral beliefs.[2] These researchers found that considering the Footbridge dilemma -- a moral dilemma that tends to elicit deontological judgments -- led people to lower their credence in the utilitarian belief, "Always take whatever means necessary to save the most lives."[3] They concluded that people revise their moral beliefs after exposure to a single

[2] These researchers were primarily interested in the connection between belief revision and "transfer" or "ordering effects" currently being investigated in the literature (see Schwitzgebel & Cushman, 2012; Wiegman & Waldmann, 2014 for discussions of these effects).

[3] The Footbridge dilemma asks participants to consider a situation in which they are on a footbridge, in between a runaway trolley and five workmen who will die if nothing is done. Participants are then told that the only way to save the lives of the five workmen is to push a stranger off the bridge and onto the tracks below where his large body will stop the trolley and he will die.

dilemma.

What's striking is that these researchers did not have to coax a belief change from their participants. Rather, participants seemed to recognize that their deontological judgments were inconsistent with an abstract moral principle they strongly agreed with, and updated their agreement with the principle accordingly.

The present research: Does the method of cases lead to true moral belief revision?

Although suggestive, Horne et al.'s (2013) findings leave open two important questions concerning (1) the nature of the changes in people's moral beliefs and (2) the mechanism that underlies these changes.

First, it is unclear whether the changes that Horne and colleagues (2013) observed in their participants' belief ratings are truly reflective of changes in their participants' underlying beliefs. For example, their participants may have changed their belief *reports* to appear consistent (e.g., in response to demand characteristics), without actually revising their beliefs. Another possibility is that Horne and colleagues' findings reflect simple emotional priming, or some other transient contextual factor caused by considering an emotionally evocative moral dilemma. If the changes in participants' belief reports reflect genuine changes in their moral beliefs, then they should remain relatively stable over time. In contrast, if demand characteristics or emotional priming can explain participants' reports, then any shift in belief ratings should be fleeting.

A second question concerns the mechanism underlying the hypothesized belief revision process. Does moral belief revision occur spontaneously, as a side-effect of simultaneously representing two conflicting beliefs, or does it require effortful deliberation, for example in response to explicit questions about one's beliefs? This latter possibility suggests that people will not revise their moral beliefs unless they are immediately prompted to do so. On the other hand,

if participants revise their beliefs without immediate prompting then this would suggest the process occurs more spontaneously, further cutting against the possibility that people will only change their belief reports to appear consistent. Regardless of whether belief revision in response to moral dilemmas requires prompting or is spontaneous, demonstrating that people change their beliefs in response to such dilemmas would provide an avenue for overcoming confirmation bias: Although one's beliefs may be comparatively impervious to external, contradictory evidence, perhaps they are less impervious to one another.

We conducted two experiments to investigate moral belief revision after exposure to a single dilemma. In Experiment 1, we examined how considering the Transplant dilemma affected participants' agreement with a utilitarian belief statement both immediately following exposure to the dilemma and after a six-hour delay. The findings suggested that exposure to the dilemma induced a genuine change in people's moral beliefs that remained stable over a delay, suggesting that the method of cases provides an avenue for approaching moral belief change. Then, in Experiment 2, we examined whether participants would spontaneously revise their beliefs in response to the Transplant dilemma by asking participants to rate their agreement with moral beliefs *only* after a delay.

General Methods for Experiments 1 and 2

Belief Selection

A great deal of research in moral psychology has focused on "utilitarian" moral judgment (e.g., Bartels & Pizarro, 2011; Greene, Nystrom, Engell, Darley & Cohen, 2004; Greene Sommerville, Nystrom, Darley, & Cohen, 2001). We examined one operationalization of utilitarianism previously used in the literature: "In the context of life or death situations, always

take whatever means necessary to save the most lives."[4] Existing research has found that the majority of participants agree with this moral principle, and their agreement ratings tend to be quite strong (Horne et al., 2013).

Although we assume that people's reports of their beliefs are reliable indicators of those beliefs, recent data also suggests that utilitarian principles guide people's judgments about moral dilemmas. Indeed, researchers have argued that people's judgments about these moral dilemmas do not owe purely to the idiosyncrasies of each case, but are instead influenced by general moral rules, at least some of which are utilitarian in nature (e.g., Horne et al., 2013; Lombrozo, 2009; Nichols & Mallon, 2006; Nichols et al., Draft; Royzman, Goodwin, & Leeman, 2011). For instance, Lombrozo (2009) found that participants' agreement with utilitarian principles predicted their judgments about moral dilemmas. In line with this finding, several studies have shown that people tend to make utilitarian moral judgments--choosing to sacrifice the life of one person to save many--across a number of different moral dilemmas (e.g., Côté, Piff, & Willer, 2013; Greene et al., 2001; 2004; Cushman, Young, and Hauser, 2006; Hauser, 2006; Moore, Clark, and Kane, 2008). So, it does not appear that participants naively assent to this principle without appreciating at least some of its negative implications. Of course, this is not to say that people wholeheartedly endorse the doctrine of utilitarianism, as some philosophers do; it is quite likely that they also hold deontological beliefs (e.g., Nichols et al., Draft).

Dilemma Selection

Prior research suggests that people revise their moral beliefs after considering the

[4] There may be more than one reading of the utilitarian statement. We have intended for participants to take a *strong utilitarian* reading, recognizing that "whatever means necessary" includes acts like killing an innocent person. However, it is also possible that participants got a *weaker utilitarian* reading of on which these sorts of acts were not salient. Given the context of the utilitarian statement (i.e., "In life or death situations …"), we think it is quite plausible that participants imagine "whatever means necessary" include acts like killing.

Footbridge dilemma (Horne et al., 2013) -- a moral dilemma that tends to elicit deontological moral judgments (e.g., Greene et al., 2001; Greene et al., 2004). However, as many people in our target population (Amazon Mechanical Turk workers) have been exposed to this dilemma in prior studies (Bauman, McGraw, Bartels, and Warren, 2014), we chose to use the Transplant dilemma, which elicits high rates of deontological judgments (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008), but has been less widely examined.

Power Analysis

An *a priori* power analysis was conducted based on an expected effect size estimated from the analyses reported by Horne et al. (2013). A good point of comparison for the current experiments is provided by the contrast between their participants' utilitarian agreement ratings after reading the Footbridge and control dilemmas, in which they observed a large effect (d = 1.10; Cohen, 1988). A power analysis conducted in G*power (Faul & Erdfelder, 1998) revealed that at least 23 participants would be required in each experimental group in order to achieve 95% power to detect effects of this size ($\alpha = .05$). Because it was difficult to predict dropout rates in our studies, we often exceeded the sample size required for 95% power.

Experiment 1

In Experiment 1 we sought to examine whether changes in participants' moral beliefs would remain stable over time. We hypothesized that people would revise their moral beliefs after considering the Transplant dilemma. As we have discussed, in past experiments participants have almost universally agreed that it is inappropriate to harvest a single patient's organs to save the many, despite the fact that this action "maximizes the good" (Greene et al., 2008). Likewise, they have also strongly endorsed the utilitarian moral principle "In the context of life or death situations, always take whatever means necessary to save the most lives" (Horne et al., 2013).

We predicted that they would revise their agreement with this principle after making a judgment about the Transplant dilemma, because they would view their judgments about the Transplant dilemma as inconsistent with the utilitarian principle. If people genuinely changed their agreement with this utilitarian belief, then we predicted that the revised agreement rating should remain even after a delay. In contrast, if participants' revisions were no longer present after a delay, then this would suggest that their reports were due to demand characteristics, emotional priming, or other transient contextual factors. The first outcome would provide substantial evidence that moral dilemmas, and the method of cases more generally, can lead to belief revision. The second outcome may undermine this proposal.

Method

Participants

This experiment was conducted online with 195 participants (49% female; mean age of 34.9 years) recruited via the Mechanical Turk Work Distribution website between the times of 10:00 am and 11:30 am Central Standard Time. These participants were then encouraged to return later that same day between 5:00pm and 6:30pm to complete part two of the experiment. Of the 195 participants who initially completed part one, 67 returned and completed part two. Total compensation for completing both parts of the experiment was $1.50.

Design, Materials and Procedure

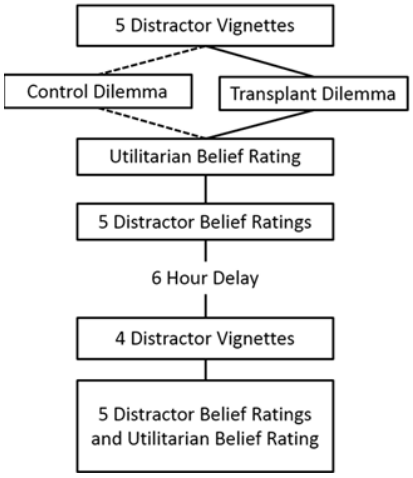

Figure 1. Diagram summarizing procedures used in Experiment 1. Participants progressed through stages of the study in order from top to bottom.

This study utilized a 2 x 2 (vignette condition x delay condition) factorial design. We manipulated the moral vignette that participants read immediately before reporting on their moral beliefs: participants were assigned to either the Control or the Transplant condition (vignette condition; between-subjects). Participants were also asked to rate their agreement with the utilitarian moral belief both immediately following their moral judgments about the Transplant dilemma and after a delay of approximately six hours (delay condition; within-subjects). Figure 1 provides a diagram summarizing the experimental procedures.

Part one of the study began with a judgment task. Participants were presented with a series of three moral and three non-moral vignettes and were asked to make judgments about these situations. In the Transplant condition, the judgment task began with five distractor vignettes and ended with a judgment about the Transplant case. In the Control condition, the judgment task ended with a non-moral control vignette, and the transplant case was replaced with a different moral vignette (a vignette about incest) in order to maintain the same number of moral and non-moral distractors. Participants made their moral judgments on a 6-point Likert

scale with endpoints labeled “Completely Inappropriate” and “Completely Appropriate.”

After the judgment task, participants completed a belief task: Participants rated their agreement with six statements expressing moral and non-moral beliefs. They made these ratings using a sliding scale, the endpoints of which were labeled “Completely Disagree” and “Completely Agree.” In order to make it difficult for participants to remember their responses to these belief statements in the second part of the experiment, there were no hashmarks or numerical indicators of their response. The only visual feedback provided by the slider was the position of the marker relative to the endpoints. The position of the marker was coded by the computer into a response between 0 and 100. Our primary dependent measure was participants’ ratings of the utilitarian belief statement “In the context of life or death situations, you should always take whatever means necessary to save the most lives.” This statement was presented at the beginning of the belief task, immediately following the moral judgment task. Five other belief statements, two-moral and two non-moral, were presented after the utilitarian belief statement. An example of a non-moral belief statement was, “I laugh out loud when someone tells me a joke that I think is funny,” and a moral belief statement was, “Incest is always morally wrong.” Additionally, half of these belief statements corresponded to the stories that participants made judgments about earlier in the experiment. The additional moral and non-moral distractors and belief statements were intended to conceal the aim of the experiment, and further reduce the likelihood of demand characteristics. After rating the six belief statements, participants completed a series of comprehension questions to ensure they paid attention to the stories that they read and then were asked to participate in part two of the experiment.

Part two was identical for participants assigned to either condition. Participants first made judgments about four unrelated distractor vignettes (two novel, two old; two moral, two non-

moral), *none* of which were the Transplant dilemma. After making judgments about these vignettes, participants again rated their agreement with five distractor belief statements and the utilitarian belief (our primary dependent measure). Three of the belief statements were originally presented in part one of the experiment and three were novel.

Results and Discussion

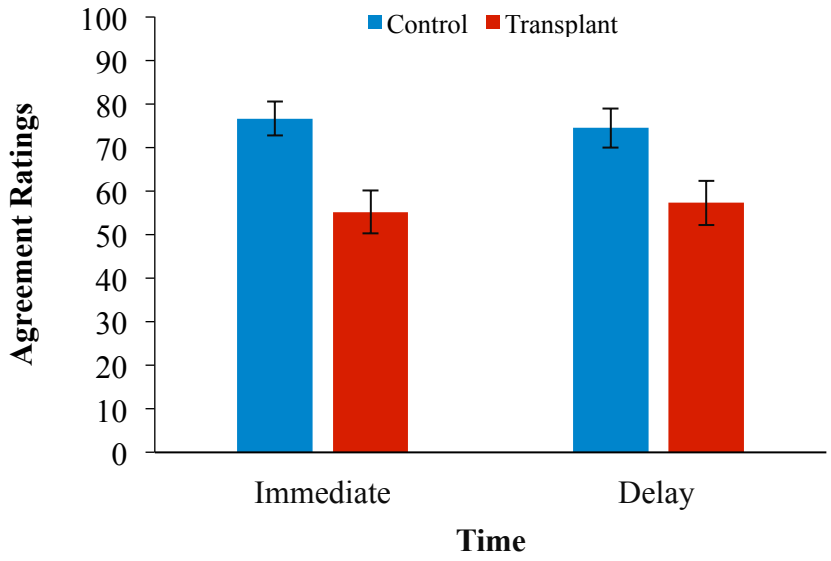

Figure 2. Average agreement ratings for the utilitarian belief statement across vignette and delay conditions in Experiment 1. Error bars represent ± 1 *SE*.

Participants' agreement ratings for the utilitarian belief statement are shown in Figure 1. Of the 195 participants who initially completed part one, 67 returned and completed part two. To rule out the possibility of self-selection effects, we compared the immediate utilitarian belief ratings of those participants who returned for part two and those who dropped out. Results of two independent samples *t*-tests revealed no reliable differences between these two groups in either the control condition ($t(99) < 1, p = .40$) or the transplant condition ($t(94) = 1.48, p = .14$).

We computed a two-way repeated measures ANOVA to determine how participants'

endorsements of the utilitarian belief were affected by reading the Transplant dilemma and by the delay. Participants in the Transplant condition gave lower ratings to the utilitarian belief than those in the Control condition, as indicated by a reliable main effect of Vignette, ($F(1, 134) = 11.32$, $p < .001$). Moreover, and consistent with our prediction that people's beliefs would remain stable over time, participants' ratings were not affected by an approximately six-hour delay $F(1, 134) < 1$, ns. Moreover, there was no reliable interaction between Vignette and Delay, $F(1, 134) < 1$, ns. Subsequent independent *t*-tests revealed that participants in the Transplant condition gave reliably lower agreement ratings for the utilitarian belief statement both immediately following the Transplant dilemma in part one, ($t(67) = 3.26$, $p < .01$, $d = .79$) and after a delay in part two ($t(67) = 2.43$, $p < .05$, $d = .59$). Both of these effects are conventionally considered medium to large effects (Cohen, 1988). Finally, a paired sample *t*-test found no reliable difference in belief ratings in the Transplant condition between parts one and two of the experiment, $t(38) = .92$, $p = .36$.

Discussion

Consistent with our predictions, the results of Experiment 1 demonstrate that people revise their beliefs after exposure to a moral dilemma and that this revision persists for at least six hours after the dilemma is considered. The stability of their revised beliefs suggests that the observed effect is not the product of emotional priming or other transient contextual factors. Rather, the results suggest that people consider their judgments about the Transplant dilemma as inconsistent with a utilitarian belief about the permissibility of killing, and revise their agreement with this utilitarian belief accordingly.

Experiment 2

Experiment 1 suggests that the method of cases may also provide a means for inducing

moral belief revision. Still, one possibility is that people will *only* revise their beliefs when they are asked about the utilitarian belief immediately after considering the Transplant dilemma. That is, it remains unclear whether merely considering the dilemma is itself sufficient to produce changes in participants' beliefs. We investigated this question in Experiment 2. This experiment was identical to Experiment 1, except that we added two conditions. In these new conditions, participants only rated their agreement with the utilitarian belief *after* the delay. If participants spontaneously revise their agreement with the utilitarian belief after considering the Transplant dilemma, then we should be able to observe the effects of this revision even when participants are not immediately prompted to rate their agreement with this belief. Experiment 2 thus served two functions: It served to directly replicate Experiment 1, and it also tested whether people will spontaneously revise their agreement with the utilitarian principle after considering the Transplant dilemma.

Methods

Participants

This experiment was conducted online with 373 participants (52% female; mean age of 36 years) recruited via the Mechanical Turk Work Distribution website. These participants were then encouraged to return later that same day to complete part two of the experiment. Of the 373 participants who initially completed part one, 120 returned and completed part two. Three participants were removed for completing the study more than once, leaving a sample of 117 participants. Total compensation for completing both parts of the experiment was $1.50.

Materials and Procedure

The stimuli and procedure of Experiment 2 were identical to those of Experiment 1, except for the inclusion of two additional conditions. In the Transplant Delay-Only and Control

Delay-Only conditions, participants rated their agreement with the utilitarian belief only after the delay. In all other respects these conditions were identical to the Transplant and Control conditions in Experiment 1. Figure 3 provides a summary diagram of the procedures in Experiment 2.

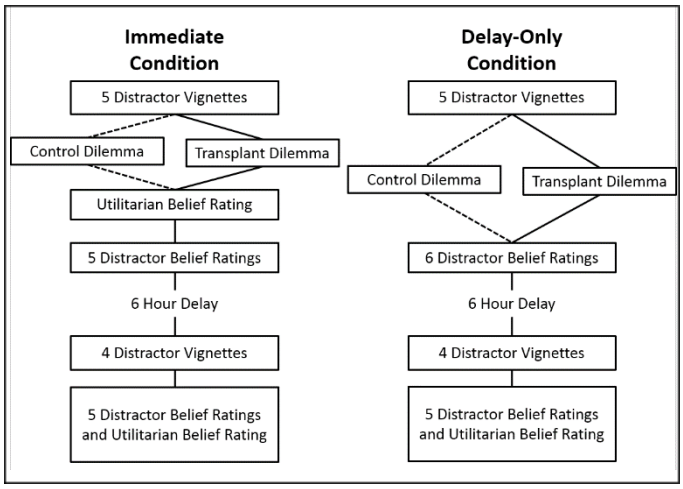

Figure 3. Diagram summarizing procedures used in Experiment 2. Participants progressed through stages of the study in order from top to bottom.

Results and Discussion

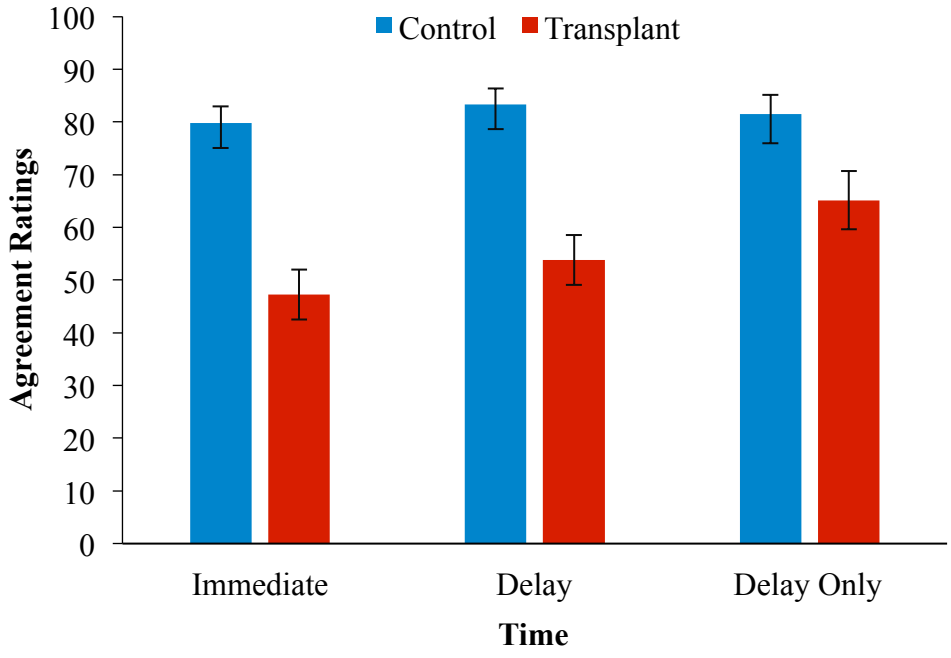

Figure 4. Average agreement ratings for the utilitarian belief statement across vignette and delay conditions in Experiment 2. Error bars represent ± 1 *SE*.

The results of Experiment 2 are shown in Figure 4. First, to rule out the possibility of self-selection effects, we again compared the immediate utilitarian belief ratings of those participants who returned for part two and those who dropped out. The results of two independent samples *t*-tests revealed no reliable differences in participants' immediate utilitarian belief ratings between these groups. This held for both the Control condition $t(96) < 1.66$, $p = .10$ and the Transplant condition $t(94) = 1.69$, $p = .09$.

The distribution of data in Experiment 2 was non-normal, rendering parametric statistics inappropriate for any subsequent hypothesis testing. Since parametric effect sizes were also inappropriate, we calculated a non-parametric common language effect size (McGraw & Wong, 1992) *CL*, as recommended by Grissom & Kim (2012). This statistic represents the probability that a score randomly drawn from one population A will be greater than a score randomly drawn from another population B.

First, we computed a series of Mann-Whitney *U* tests to determine whether Experiment 2 replicated the results of Experiment 1. Our findings replicated all previously observed effects, demonstrating again that participants revised their beliefs after reading the Transplant dilemma, and that those revisions remained stable even after a significant delay. Agreement ratings for the utilitarian belief were higher in the Control condition than in the Transplant condition when tested immediately ($U(27, 36) = 792$, $p < .001$, $CL = .815$), and after a delay ($U(27, 36) = 773.5$, $p < .001$, $CL = .796$). Participants' immediate and delayed belief ratings were examined with a related-samples Wilcoxon Signed Rank test. The ratings did not reliably differ over time in the Transplant condition ($W(26) = 430.5$), $p = .058$), although this effect approached significance.

However, a similar trend was observed in the Control condition ($W(27) = 200$, $p = .059$), suggesting that this trend was not associated with our experimental manipulation.

Of particular interest in Experiment 2 was whether people would spontaneously recruit and revise their beliefs after considering the Transplant dilemma—that is, whether participants' agreement ratings with the utilitarian belief would be lower in the Delay-Only Transplant condition than in the Delay-Only Control condition. This test revealed a significant effect, where participants in the Delay-Only Transplant condition gave lower agreement ratings for the utilitarian belief than participants in the Delay-Only Control condition, $U(27, 26) = 471.5$, $p = .032$, $CL = .673$. We also compared participants' delayed utilitarian agreement ratings from the Immediate-Delay Transplant condition and the Delay-Only Transplant condition. There were no reliable differences in their agreement ratings ($U(36, 26) = 360$, $p = .12$), even though participants in the Immediate-Delay Transplant condition also rated their agreement with the utilitarian belief immediately after considering the Transplant dilemma earlier that day. This result is particularly striking as it suggests that people spontaneously recruit and revise their beliefs when considering moral dilemmas.

General Discussion

How is it that moral convictions are at once steadfast, and yet people can and do change their moral beliefs? And what factors can lead to authentic moral belief change? The present research demonstrates that considering a moral dilemma can produce authentic change in people's moral beliefs. We found that making a judgment about the Transplant dilemma changed people's beliefs in the utilitarian principle, "In the context of life or death situations, always take whatever means necessary to save the most lives." This was the case both immediately following exposure to the dilemma and six hours later. Experiment 2 demonstrated that this change occurs

even when participants are not immediately queried about their agreement with the utilitarian principle. This latter result suggests that being exposed to moral dilemmas may lead people to revise their beliefs, even when they are not prompted to do so.

It is unlikely that participants' belief reports owe simply to a desire to appear consistent, as our experiments included distractor vignettes and distractor belief statements that concealed our aim. Additionally, research on analogical problem solving has shown that, in the absence of featural support for memory retrieval, it is rare for participants to recall a relevant problem when attempting to solve a new problem, even after a short delay (e.g., Gick & Holyoak, 1983; Wharton & Holyoak, 1996). Accordingly, it is unlikely that participants in Experiment 2 recalled the Transplant dilemma when they rated their agreement with the utilitarian belief six hours later. Instead, a more plausible interpretation of the results of Experiment 2 is that exposure to the Transplant dilemma changed participants' utilitarian beliefs when they first encountered the dilemma. These changes then persisted, affecting the participants' agreement ratings six hours later. These results suggest that participants exhibit authentic revisions in their moral beliefs after considering a single moral dilemma.

What are the mechanisms underlying moral belief revision?

Our results suggest that moral dilemmas, which pit competing moral beliefs against each other, can lead people to revise their beliefs. This marks a significant advance for understanding the psychological processes underlying moral conviction and confirmation bias. It seems that people possess many moral beliefs that can be mapped onto a specific moral situation (e.g., Nichols & Mallon, 2006; Prinz, 2007). Once such a mapping is computed, a belief may suggest a certain response to the moral situation under consideration. When thinking about a moral dilemma, people might successfully map multiple beliefs to the same situation. If different

beliefs entail different responses to the same situation, then this suggests that the elicited beliefs are inconsistent. In this case, one or more beliefs must be revised in order to restore coherence. For example, our participants may have reduced their agreement with the utilitarian belief after considering the Transplant dilemma because they mapped a utilitarian belief to the dilemma but ultimately recognized that their judgment about the appropriate action in the dilemma was inconsistent this belief.

Of course, there are likely a number of factors that could interfere with this type of mapping, such as cognitive load or increased complexity. Additionally, if convictions are held strongly enough, it is possible that people might respond to perceived inconsistency by rejecting the mapping between a case and their belief. This might be more likely to be the case for convictions that have been publically expressed, such as about political or social issues, which have largely been the focus of research on moral conviction (e.g., Skitka, 2010). Further research is necessary to more completely understand the connection between people's judgments about dilemmas and moral belief revision.

Conclusion

The present studies reveal a means for inducing moral belief revision. Rather than presenting people with data that directly contradict their existing moral beliefs—and which could therefore be ignored or discounted due to confirmation bias—moral dilemmas, and the method of cases more generally, may circumvent confirmation bias by highlighting existing (but perhaps heretofore unnoticed) inconsistencies in people's beliefs. While moral belief revision is a necessary first step toward moral progress, contemporary moral psychologists have often seen it as insurmountable. This study gives reason for optimism, as we have provided a method for inducing moral belief revision.

References

Bauman, C. W., McGraw, A. P., Bartels, D. M., & Warren, C. (forthcoming). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in Moral Psychology. *Personality and Social Psychology Compass*.

Cooper, J. (2007). *Cognitive dissonance: 50 years of a classic theory*. London: SAGE.

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.

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*(1), 1-38.

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *The Neuroscience of morality: Emotion, brain disorders, and development* (pp. 35-80). MIT Press

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*(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.

Grissom, R., & Kim, J. (2012). *Effect sizes for research: Univariate and multivariate applications*. New York: Routledge.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814-834.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557-559.

Hauser, M. D. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Ecco.

Horne, Z., Powell, D., & Spino, J. (2013). Belief updating in moral dilemmas. *Review of Philosophy and Psychology*, *4*(4), 705-714.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220.

Kamm, F.M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

Klayman, J. (1995). Varieties of Confirmation Bias. *Psychology of Learning and Motivation - Advances in Research and Theory*, *32*(C), 385-418.

Kohlberg, L., & Hersh, R. H. (1977). Moral development: A review of the theory. *Theory into Practice*, *16*(2), 53-59.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science*, *33*(2), 273-286.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098-2109.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361-365.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science*, *19*(6), 549-557.

Nagel, J. (2012). Intuitions and experiments: A defense of the case method in epistemology. *Philosophy and Phenomenological Research*, *85*(3), 495-527.

Nichols, S. (2004). *Sentimental Rules: On the natural foundations of moral judgment*. New York: Oxford University Press.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition*, *100*(3), 530-542.

Nichols, S., Kumar, S., & Lopez, T. (Draft). Rational learners and non-utilitarian learners.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175-220.

Parfit, D. (2013). *On what matters*. New York: Oxford University Press

Prinz, J. (2007). *The Emotional Construction of Morals*. New York: Oxford University Press.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57-75.

Rochat, P., Dias, M. D., Liping, G., Broesch, T., Passos-Ferreira, C., Winning, A., & Berg, B. (2009). Fairness in distributive justice by 3-and 5-year-olds across seven cultures. *Journal of Cross-Cultural Psychology*, *40*(3), 416-442.

Royzman, E. B., Goodwin, G. P., & Leeman, R. F. (2011). When sentimental rules collide: "Norms with feelings" in the dilemmatic context. *Cognition*, *121*(1), 101-114.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*(2), 135-153.

Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass*, *4*(4), 267-281.

Sloane, S., Baillargeon, R., and Premack, D. (2012). Do Infants Have a Sense of Fairness? *Psychological Science*, *23*(2), 196-204.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204-217.

Turiel, E. (2006). The development of morality. In W. Damon & R. M. Lerner (Series Eds.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 789–857). New York: Wiley.

Wharton, C. M., Holyoak, K. J., & Lange, T. E. (1996). Remote analogical reminding. *Memory & Cognition*, *24*(5), 629-643.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28-43.