Belief Updating in Moral Dilemmas

Research on two classic moral dilemmas, the Trolley and the Footbridge dilemma, suggests that one's past moral experiences can affect one's subsequent moral decisions. These dilemmas have interested moral psychologists, in part, because they have found that people's judgments about the dilemmas are affected by the order in which the dilemmas are considered. Furthermore, this effect is asymmetrical: people who consider the Trolley dilemma after the Footbridge dilemma make significantly different judgments than people in control conditions, but the converse is not true. We argue that this asymmetry is the result of a difference in how the each dilemma affects pre-existing beliefs regarding the importance of saving lives. In two experiments, we show that considering the Footbridge dilemma disconfirms these pre-existing beliefs, while considering the Trolley dilemma does not significantly affect them. Consistent with predictions of belief updating models of ordering effects, these findings offer a clear and parsimonious account of the asymmetry.

Can the moral decisions you have made in the past affect your thinking about current moral situations? Research on two classic moral dilemmas, the Trolley and Footbridge dilemmas, suggests that they can. In the last twenty years, these dilemmas have interested researchers, in part, because people's judgments about the dilemmas are affected by the order in which the dilemmas are considered. Moreover, this order effect is asymmetrical: if people consider the Trolley dilemma after making a judgment about the Footbridge dilemma, they are significantly less likely to deem it acceptable to redirect the train, but reading the Trolley dilemma has little to no effect on subsequent Footbridge judgments (Cushman & Schwitzgebel, 2012). We will refer to this finding as the *Footbridge-Trolley ordering effect*.

Psychologists and philosophers alike have discussed this finding (Feltz & Cokeley, 2011; Graham, In Press; Liao et al., 2011; Lombrozo, 2010; Petronovich & O'Neil, 1996; Sinnott-Armstrong, 2007; Wiegmann et al., 2011; Wiegmann & Okan, 2012). Some have argued that the ordering effect provides critical insight into the mental processes that underlie people's moral judgments (Cushman & Schwitzgebel, 2012), while others have interpreted this finding as substantial evidence against ethical theories or positions in moral epistemology (Liao et al.,

2011; Sinnott-Armstrong, 2007). Despite extensive research and philosophical discussion of this effect, there is no consensus explanation as to how or why it occurs.

The explanations that have been proposed for ordering effects in moral judgment have often been incomplete. For example, some researchers suggest that ordering effects in moral judgment are produced when participants compare similar moral vignettes (Liao et al., 2011; Petronovich & O'Neil, 1996; Sinnott-Armstrong, 2007; Wiegmann et al., 2011). Although it is likely that a process of comparison contributes to ordering effects, this suggestion does not offer an explanation of why the effect is asymmetrical.

Cushman and Schwitzgebel (2012) proposed that the asymmetry is caused by a difference in the mental processes that are recruited when people make utilitarian or deontological moral judgments. These researchers assume a dual-process account of moral reasoning wherein different mental processes produce different types of moral judgments (Greene et al., 2001). Specifically, it has been proposed that emotional processing leads to deontological moral judgment whereas reasoning processes lead to utilitarian moral judgments. Based on this account, Cushman and Schwitzgebel argue that the emotional processes typically associated with deontological moral judgment are not sensitive to new evidence, such as a recently considered moral case. In contrast, the reasoning processes typically associated with utilitarian moral judgment may be responsive to new evidence. Thus, differences in processing explain why people's judgments about the Trolley dilemma (utilitarian) are influenced by their previous judgment about the Footbridge dilemma (deontological), but the opposite is not true.

As noted, this account presumes that the Trolley and Footbridge dilemmas recruit fundamentally different mental processes. The asymmetry in the ordering effect may be construed as further evidence that different mental processes underlie deontological and

utilitarian judgment. However, asymmetrical ordering effects have been widely reported in research on judgments of non-moral problems, and have been explained without appeal to differences in mental processes (Hogarth & Einhorn, 1992). Moreover, some researchers have found evidence that moral judgments recruit domain general cognitive processes such as decision-making heuristics (Rai & Holyoak, 2010), and mechanisms associated with causal reasoning (Waldmann, 2012). In contrast to Cushman and Schwitzgebel's proposal, these researchers argue that the field should seek domain-general explanations of moral behavior before positing more specialized mental processes that appeal specifically to the details of the dilemmas under consideration.

In the present paper we sought to examine whether domain general models of learning and judgment can provide insight into the Footbridge-Trolley ordering effect. Sequential learning models have been developed across a variety of cognitive domains, describing the formation of stimulus-response associations in classical conditioning (Rescorla & Wagner, 1972), learning about causal relationships (Shanks, 1985; Van Hamme & Wasserman, 1994; Danks et al., 2003) and more general cases of belief-updating (Hogarth & Einhorn, 1992). These models describe how the state of some mental representation is affected by new experiences. Importantly, these mental representations are bounded, i.e., they have maximum and minimum states, often expressed as probabilities of 0 and 1 (Danks et al., 2003; Hogarth & Einhorn, 1992). As a general principle in models of sequential updating, the degree to which some experience affects a given mental representation is contingent on the current state of that representation: the effect will be inversely proportional to the difference from the representation’s current state to whichever bound it is approaching. It follows that when a belief or association is strong, experiences supporting the representation are less influential than experiences that undermine it,

whereas just the opposite is true when the representation is weak. Thus, if the Footbridge-Trolley ordering effect is the result of a process of sequential learning or belief updating, then this weighting dynamic might explain the effect's asymmetry.

We argue that people's judgments about the Footbridge and Trolley dilemmas are influenced by the state of their general underlying moral beliefs, and that considering these moral dilemmas provides evidence for or against these beliefs. We focus our examination on Hogarth and Einhorn's (1992) Belief-Adjustment model, a sequential learning model that was explicitly developed to explain ordering effects that occur when individual pieces of evidence affect existing beliefs. However, as discussed above, the general principles we describe are at work in a number of models that describe a wide range of human (and nonhuman animal) behavior (Danks et al., 2003; Hogarth & Einhorn, 1992; Rescorla & Wagner, 1972; Shanks, 1985; Van Hamme & Wasserman, 1994).

A Belief Adjustment Explanation

Hogarth and Einhorn's belief-adjustment model (1992) specifies how the state of some pre-existing belief is affected by new evidence. The following equation describes the model:

$$S_k = S_{k-1} + w_k[s(x_k) - R].$$

S_k is the state of some belief after k pieces of evidence and S_{k-1} is the state of that belief just before the k^{th} piece of evidence. Intuitively, 0 might describe a belief state like 'complete disbelief', and 1 'complete belief.' The variable $s(x_k)$ is the individual's subjective evaluation of the k^{th} piece of evidence. The variable R is a reference point against which the k^{th} piece of evidence is evaluated and w_k is the weight given to this contrast. Each of these terms takes on a value between 0 and 1.

To help frame the Footbridge-Trolley ordering effect in the context of prior research on ordering effects, we must consider what belief, if any, is being adjusted when people consider these moral dilemmas. For naive participants, S_{k-1} should not be construed as being about either the Trolley or the Footbridge dilemmas in particular, since participants have never considered either dilemma prior to participating in the experiment. Rather, S_{k-1} would be a more general moral belief recruited when one considers these moral situations. For example, a candidate belief might be "In life or death situations, one should always take whatever means necessary to save the most lives." In this case, participants' judgments about a dilemma constitute their evaluation of a new piece of evidence $s(x_k)$ that is relevant to their more general belief S_{k-1}. Hogarth and Einhorn's model describes how these evaluations affect S_{k-1}, resulting in the updated belief S_k, the belief recruited when making moral judgments about subsequent similar dilemmas (Petrinovich & Oneill, 1996). One standing question is, why do the Footbridge and Trolley dilemmas affect the general pre-existing belief S_{k-1} to different degrees?[1]

Like other sequential learning models, in the belief-adjustment model (Hogarth & Einhorn, 1992) the role of evidence is affected by the current state of the belief, *S*. Under the belief- adjustment model, when evidence is disconfirmatory the weight is given as

$$w_k = \alpha S_{k-1},$$

whereas if evidence is confirmatory, the weight is given as

$$w_k = \beta(1-S_{k-1}).$$

[1] There is another question we might ask along these lines: What beliefs other than those described by *S* might affect either Footbridge or Trolley judgments alone? For example, it is possible that some additional background beliefs might influence Footbridge Judgments but play no role in Trolley judgments (e.g., "it is wrong to use someone as a means to an end"). In this case, making a judgment about the Trolley dilemma might influence *S*, but would have no effect on these other beliefs. Thus, the Trolley dilemma might have a relatively weaker effect on Footbridge judgments, as it affects only some of the beliefs important to Footbridge judgments.

The weight of disconfirmatory evidence is proportional to the strength of the pre-existing belief (S_{k-1}), whereas the weight of confirmatory evidence is proportional to its inverse ($1 - S_{k-1}$). The parameters α and β ($0 < \alpha, \beta < 1$) represent an individual's sensitivity to disconfirmatory and confirmatory evidence, respectively. We will assume they are equal and set them equal to 1 for simplicity.

For the sake of argument, suppose that participants' pre-existing belief, S_{k-1}, represents a proposition like "In life or death situations, one should always take whatever means necessary to save the most lives." If this is the case, then it is reasonable to believe that people would not be indifferent about the truth of this belief, and in fact might endorse it strongly. That is, if we assign it a value in the model, it would be greater than .5. Under these conditions, the belief-adjustment model predicts that disconfirmatory evidence (such as the judgments the Footbridge dilemma tends to elicit) will have a stronger effect on this belief S_{k-1} than equivalent confirmatory evidence (like the judgments the Trolley dilemma tends to elicit). As a toy example, suppose S_{k-1} takes on a value of .75. The weight of disconfirmatory evidence is $1 \times .75 = .75$, whereas the weight of confirmatory evidence is $1 \times (1 - .75) = .25$. Suppose further that the Trolley and Footbridge dilemmas provide equivalent evidence for and against the belief, such that the contrast $[s(x_k) - R]$ is equal to .5 in both cases. Assuming these values, considering the Trolley dilemma will increase S_{k-1} by .125 ($S_k = S_{k-1} + .125$), whereas considering the Footbridge dilemma will decrease S_{k-1} by .375 ($S_k = S_{k-1} - .375$). This is merely an illustration, but whenever we assume some value for S_{k-1} greater than .5, the weight for disconfirmatory evidence will be greater than the weight for confirmatory evidence (all else being equal).

In sum, if the beliefs that undergird people's judgments about the Footbridge and Trolley dilemmas are held strongly, then the asymmetry in the Footbridge-Trolley ordering effect is a

natural consequence of a sequential learning model such as Hogarth and Einhorn's (1992) belief-adjustment model. In two experiments, we show how the Footbridge and Trolley dilemmas affect people's pre-existing beliefs about morality and examine how the belief-adjustment model (1992) can explain the asymmetry in the Footbridge-Trolley ordering effect.

Experimental Design

We sought to investigate how consideration of the Footbridge and Trolley dilemmas affects participants' pre-existing moral beliefs and the degree to which they endorse these moral beliefs. To this end, we conducted two experiments. In each experiment, participants made judgments about either the Footbridge dilemma, the Trolley dilemma, or a non-moral control dilemma (between subjects design) and then rated the extent to which they agree with a general moral belief statement. People's responses to moral dilemmas are often discussed in terms of "utilitarianism" and "deontology" (Greene et al., 2001; Nichols & Mallon, 2006; Prinz, 2007; Wiegmann & Okan, 2012). Accordingly, in Experiment 1 participants rated a statement generally representing a utilitarian viewpoint: "In life or death situations, you should always take whatever means necessary to save the most lives." In Experiment 2, participants rated a statement that represented a more deontological viewpoint: "You should never kill another person." Participants' agreement ratings in control conditions were used to measure their endorsement of the utilitarian and deontological beliefs described above. By comparing the agreement ratings of participants who made moral judgments about the Footbridge and Trolley dilemmas with ratings of participants' in the control condition, we were able to measure the degree to which these dilemmas confirm or disconfirm the pre-existing beliefs researchers have hypothesized are recruited when making judgments about the Trolley and Footbridge dilemmas.

First, we predicted both the utilitarian and deontological belief would be held strongly. Second, we predicted that the Footbridge dilemma would be stronger disconfirmatory evidence for utilitarian belief than the Trolley dilemma would be confirmatory evidence for that belief. This result would explain the asymmetry in the Footbridge-Trolley ordering effect. However, it is also possible that the Trolley dilemma affects other relevant beliefs more strongly than the Footbridge dilemma, which could undermine this explanation. To begin to rule this out, in Experiment 2 we examined the effects of these dilemmas on a deontological belief.

Experiment 1

Method

This experiment was conducted online, administered using Qualtrics survey software. To eliminate the possibility of automated computer responses, participants were required to pass a CAPTCHA test at the beginning of the experiment. After reading instructions, participants first read one dilemma and made a judgment about the appropriate action. Participants indicated their judgment about the moral dilemma using a 6-point Likert scale with the endpoints labeled as "Completely Inappropriate" to "Completely Appropriate." Then, they rated the extent to which they agree with the general moral belief statement. In Experiment 1, this was: "In life or death situations, one should always take whatever means necessary to save the most lives" (utilitarian belief statement). Participants rated their agreement using a 7-point Likert scale with endpoints labeled as "Completely Disagree" to "Completely Agree." Afterwards, participants answered a pair of simple reading comprehension questions about the dilemma they read, and also indicated whether they had ever seen or heard the dilemma before. Participants advanced through the experiments at their own pace, but timing controls ensured they stayed on each page long enough to completely read instructions, dilemmas, and questions.

Participants

In Experiment 1, 265 participants (97 female, mean age = 30 years) were recruited online, through the Amazon Mechanical Turk (mTurk) work-distribution website. To be eligible for the mTurk posting, workers had to reside in the U.S. and have at least a 95% approval rate. Participants were compensated $.20 for their participation.

Results and Discussion

Participants who had previously seen the dilemma they read (52)[2] or failed the reading comprehension checks (an additional 24) were excluded from analysis. This left 189 participants in the final analysis.

Consistent with our hypothesis, participants in the control condition strongly endorsed the utilitarian belief statement (mean rating = 5.62). Endorsement was also high in the Trolley condition (mean = 5.26) and there were no significant differences between ratings in the Trolley and control conditions, $t(137) = 1.657$, $p = .07$. In contrast, participants who read the Footbridge dilemma rated their agreement with the statement significantly lower (mean = 4.10), $t(137) = -5.79$, $p < .001$. Moreover, participants in the Footbridge condition gave significantly lower agreement ratings than participants in the Trolley condition, $t(98) = -4.278$, $p < .001$. These results suggest that people take the Footbridge dilemma as evidence against this utilitarian belief statement, but do not see the Trolley as evidence in favor of it (in fact, the trend in Trolley condition was exactly the opposite).

Participants' moral belief agreement ratings significantly correlated with their moral judgments in both the Footbridge condition ($r(48) = .326$, $p < .05$) and in the Trolley condition

[2] This number is fairly large, but it is consistent with the proportion of respondents who had previously been exposed to these cases as reported by Graham (2012?).

($r(48) = .459$, $p < .01$), suggesting that this moral belief is importantly related to these moral judgments.

Experiment 2

Method

The design and materials of Experiment 2 were identical to those of Experiment 1, with the exception of the background belief statement. In Experiment 2, participants rated their agreement with the statement: "You should never kill another person" (deontological belief statement). In Experiment 2, 202 participants (85 female, mean age = 31 years) were recruited via mTurk using the same criteria and with the same compensation as in Experiment 1.

Results and Discussion

Again, participants who had previously seen the dilemma they read (40) or failed the reading comprehension checks (21) were excluded from analysis. This left 141 participants in the final analysis.

Participants in all three conditions strongly endorsed the deontological belief statement. There were no significant differences between the control (mean = 5.08) and Footbridge conditions (mean = 5.43, $t(94) = 0.972$, $p = .33$), nor between the control and Trolley conditions (mean = 5.49, $t(102) = 1.168$, $p = .25$). Participants' moral belief agreement ratings correlated only weakly (and non-significantly) with their moral judgments in both the Footbridge condition ($r(35) = -.21$, $p = .21$) and in the Trolley condition ($r(43) = -.29$, $p = .06$). It appears that neither the Footbridge nor the Trolley dilemmas are taken as evidence for or against this deontological belief statement, and that the strength of this belief may not be as relevant for the formation of judgments about these dilemmas.

General Discussion

We examined how two pre-existing moral beliefs, one representing a utilitarian perspective and the other representing a deontological perspective, play a role in producing the asymmetrical Footbridge-Trolley ordering effect. We were interested in three factors: how strongly the utilitarian and deontological beliefs are endorsed, how they relate to people's moral judgments in the Footbridge and Trolley dilemmas, and how consideration of these dilemmas might affect people's endorsement of these moral beliefs. We found that people strongly endorsed both beliefs that we examined, but that the utilitarian belief was more strongly related to their moral judgments. Furthermore, we found that considering the Footbridge dilemma disconfirmed the utilitarian belief, as we predicted, but the Trolley dilemma did not affect it either way. Neither of these dilemmas confirmed or disconfirmed the deontological belief. We propose that this pattern of results explains the asymmetry in the Footbridge-Trolley ordering effect and is consistent with sequential learning and belief adjustment models like Hogarth and Einhorn's belief adjustment model of ordering effects (1992).

If we consider the utilitarian belief to as S_{k-1}, then we must conclude that $w_{\text{Footbridge}}[s(x_{\text{Footbridge}}) - R]$ is not equal to $w_{\text{Trolley}}[s(x_{\text{Trolley}}) - R]$. Thus, the two possible sources of this inequality are the subjective evaluations of the dilemmas ($s(x_k)$) and the weights (w_k) given to those evaluations. When a to-be-adjusted belief is strong, as is the case with the utilitarian belief in question, the belief adjustment model predicts that disconfirmatory evidence will be weighted strongly, whereas confirmatory evidence will be weighted weakly. This is precisely what we observed. Disconfirmatory evidence in the form of the Footbridge dilemma had a strong effect, whereas the Trolley dilemma had no significant effect. Although it is possible that the dilemmas are also given different evaluations (after all, they are similar but not identical), this difference in weights is sufficient to explain our findings regarding the utilitarian

belief. In this way, the belief adjustment model (Hogarth & Einhorn, 1992) provides a domain-general explanation for the Footbridge-Trolley ordering effect.

Cushman and Schwitzgebel (2012) have argued that the Footbridge-Trolley ordering effect is caused by a process of comparison and that the asymmetry in the ordering effect can be attributed to the different mental processes recruited when people make judgments about these dilemmas. However, Cushman and Schwitzgebel's explanation appears theoretically unnecessary to explain the asymmetry in the Footbridge-Trolley ordering effect. Rather, as predicted by the belief adjustment model, which also explains *many* other asymmetrical ordering effects (Hogarth & Einhorn, 1992), the Footbridge and Trolley dilemmas disconfirm the pre-existing moral beliefs recruited when considering these dilemmas to different degrees. As Hogarth and Einhorn (1992) predict, a difference in how evidence affects one's credence in these pre-existing moral beliefs, and the strength of one's credence in those beliefs, can both result in an asymmetrical ordering effect.

Directions for Future Research

These experiments do leave some questions unanswered. For one, it is not entirely clear why consideration of the Trolley dilemma does not affect the deontological belief. After all, since we found that the deontological belief was held strongly, the belief adjustment model (Hogarth & Einhorn, 1992) would predict that disconfirmatory evidence should be weighted strongly. One possible explanation is that people simply do not view their judgments about the Trolley dilemma as strong evidence against deontological principles. In line with this, we found that endorsement of the deontological belief statement did not strongly predict participants' Trolley judgments. If this belief is not relevant to judgments about the dilemma, then it seems unlikely that it would be strongly influenced by those judgments.

Another issue that requires further investigation is what other moral beliefs are adjusted by consideration of these moral dilemmas. We examined how the Footbridge and Trolley dilemmas affect just two specific moral beliefs, chosen on the basis of prior research that suggested they were likely to be related to Footbridge and Trolley judgments (Nichols & Mallon, 2006; Prinz, 2007; Wiegmann & Okan, 2012). Consistent with this hypothesis, we found that the utilitarian belief statement in particular accounted for a significant proportion of the variance in people's Footbridge and Trolley judgments. However, the majority of the variance in people's judgments remains unexplained. Clearly, there are many other beliefs that are relevant to judgments about these moral dilemmas, and that might also be influenced by consideration of them.

Despite these limitations, the results of our experiment are encouraging as they show how mathematical models can explain belief updating in the domain of moral judgment. More generally, these results provide an example of how domain-general decision making models might be used to make predictions about moral cognition.

References

Cushman, F., & Schwitzgebel, E. (In Press). Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers. *Mind & Language*

Danks, D., Griffiths, T. L., & Tenenbaum, J. B. (2003). Dynamical Causal Learning. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems* (pp. 67–74). Cambridge, Mass: The MIT Press.

Feltz, A., & Cokely, E. T. (2011). Individual differences in theory-of-mind judgments: Order effects and side effects. *Philosophical Psychology*, *24*(3), 343-355.

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.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A Dissociation Between Moral Judgments and Justifications. *Mind Language*, *22*(1), 1-21.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1-55.

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2011). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, 1-11.

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

Neill, P. O. (1996). Original Articles Influence of Wording and Framing Effects on Moral Intuitions. *Ethology and Sociobiology*, *171*(3), 145-171.

Prinz, J. (2007). *The Emotional Construction of Morals*, OUP.

Rai, T. S., & Holyoak, K. J. (2010). Moral Principles or Consumer Preferences? Alternative Framings of the Trolley Problem. *Cognitive Science*, *34*(2), 311-321.

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy (Eds.), *Classical Conditioning II*, pp. 64–99. New York: Appleton-Century-Crofts.

Shanks, D. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology Section B: Comparative and Physiological Psychology*, 37(1), 1–21.

Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In W. Sinnott-Armstrong (Ed.), *Moral Psychology Volume 2 The Cognitive Science of Morality Intuition and Diversity* (pp. 47-76). MIT Press.

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp. *Philosophy and Phenomenological Research*, *76*(1), 138-155.

Van Hamme, L.J. & Wasserman, E.A. (1994). Cue competition in causality judgements: The role of nonpresentation of compound stimulus elements.*Learning and Motivation*, 25, 127–151.

Waldmann, M.R., Nagel, J., and Weigmann, A. (2012). Moral Judgment. (K. J. Holyoak & R. G. Morrison, Eds.) *The Oxford handbook of thinking and reasoning*, 1-68.

Wiegmann, A., Nagel, J., & Mangold, S. (2008). Order Effects in Moral Judgment. *Philosophical Psychology*, (March), 2111-2116

Wiegmann, A., & Okan, Y. (2012). Order Effects in Moral Judgment: Searching for an Explanation. *Proceedings of the Thirty-Fourth Annual Conference of the Cognitive Science Society.* Sapporo, Japan.

Ziefle, M. (1998), Effects of display resolution on visual performance, Human Factors, 40(4), 555–568.