COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 163–177

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2011.01210.x

Reflection and Reasoning in Moral Judgment

Joseph M. Paxton,[a] Leo Ungar,[b] Joshua D. Greene[a]

[a]*Department of Psychology, Harvard University*
[b]*Stanford University School of Medicine*

Received 4 November 2010; received in revised form 4 March 2011; accepted 18 April 2011

Abstract

While there is much evidence for the influence of automatic emotional responses on moral judgment, the roles of reflection and reasoning remain uncertain. In Experiment 1, we induced subjects to be more reflective by completing the Cognitive Reflection Test (CRT) prior to responding to moral dilemmas. This manipulation increased utilitarian responding, as individuals who reflected more on the CRT made more utilitarian judgments. A follow-up study suggested that trait reflectiveness is also associated with increased utilitarian judgment. In Experiment 2, subjects considered a scenario involving incest between consenting adult siblings, a scenario known for eliciting emotionally driven condemnation that resists reasoned persuasion. Here, we manipulated two factors related to moral reasoning: argument strength and deliberation time. These factors interacted in a manner consistent with moral reasoning: A strong argument defending the incestuous behavior was more persuasive than a weak argument, but only when increased deliberation time encouraged subjects to reflect.

Keywords: Dual-process model; Moral decision making; Moral judgment; Moral psychology; Moral reasoning; Morality; Reflection; Social intuitionist model

Can reflecting on a moral question change one's mind? Are people amenable to moral reasoning? For decades, the obvious answers were ''yes'' and ''yes'' (Kohlberg, 1969; Turiel, 1983). Since Haidt's (2001) influential critique of rationalist moral psychology, the roles of reflection and reasoning in moral judgment have remained unclear. While many researchers, including Haidt himself (2001, 2007; Haidt & Kesebir, 2010), believe that reflection and reasoning play significant roles in moral judgment, the evidence for this claim remains surprisingly limited (Paxton & Greene, 2010). Our present aim is to document and characterize the influence of reflection and reasoning on moral judgment.

Correspondence should be sent to Joseph M. Paxton, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138. E-mail: jpaxton@wjh.harvard.edu

According to Haidt's Social Intuitionist Model (SIM), moral judgment is predominantly intuitive, driven primarily by automatic emotional responses that are effortless and produced by unconscious processes. Numerous studies now attest to the influence of automatic emotional responses on moral judgment (e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Koenigs et al., 2007; Schnall, Benton, & Harvey, 2008; Valdesolo & DeSteno, 2006; Wheatley & Haidt, 2005). As noted above, the roles of reflection and reasoning—here understood as processes that are conscious, controlled, and often temporally extended (Haidt, 2001)—remain far less clear.

According to the SIM, reflection and reasoning typically serve to rationalize moral judgments that were previously made intuitively. The SIM also posits that one may occasionally influence one's own judgments directly through reasoning or indirectly through the influence of one's reasoning on one's intuitions. Finally, the SIM allows that one's expressed reasoning can influence another's judgment by influencing that person's intuitions. Other theories give moral reflection and reasoning more prominent billing, at least rhetorically, and perhaps more substantively. For example, Greene et al.'s (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001; Greene, 2012) dual-process theory is consistent with the ubiquity of utilitarian (cost–benefit) moral reasoning. Pizarro and Bloom (2003) hypothesize that moral reasoning plays a prominent role in shaping moral intuitions, and Nichols (2004) emphasizes the influence of moral rules that are grounded in emotion but applied via reasoning. We are not at present concerned with whether and to what extent different theories of moral psychology can ''claim'' moral reflection and reasoning. Our concern is with the state of the evidence.

Paxton and Greene (2010) recently reviewed the evidence for the influence of moral reasoning and concluded that the evidence is suggestive but limited. Greene et al. have implicated controlled cognitive processes in moral judgment using fMRI (2001, 2004) and reaction time data (2008), while Bartels (2008) and Moore, Clark, and Kane (2008) have produced consistent results by examining individual differences. Controlled regulatory processes have also been implicated in morally questionable behavior, such as rationalizing moral hypocrisy (Valdesolo & DeSteno, 2007). However, the nature of these controlled processes remains unclear. More specifically, one might argue that these controlled processes merely manage competing intuitions and do not involve genuine moral reasoning. People modify their judgments when explicitly instructed to make ''rational, objective'' judgments (Pizarro, Uhlmann, & Bloom, 2003), reject judgments that are inconsistent with other judgments when the inconsistency is highlighted by experimenters (Cushman, Young, & Hauser, 2006; Paharia, Kassam, Greene, & Bazerman, 2009), and appear to override certain negative implicit attitudes (Inbar, Pizarro, Knobe, & Bloom, 2009). There is evidence that attitudes can, in general, be modified by reasoned arguments (Petty & Cacioppo, 1986), but moral attitudes may be particularly stubborn.

Thus, past research hints at the efficacy of moral reflection and reasoning but falls far short of answering the question with the greatest social significance (Bloom, 2010; Greene, 2012): Can reflection and reasoning make people change their moral views, and, if so, under what circumstances? The present research addresses this question in two ways. In Experiment 1, we ask whether exposing people to non-moral problems with counter-intuitive

answers can induce them to make counter-intuitive moral judgments. In Experiment 2, we ask whether giving people a strong, counter-intuitive argument, coupled with time to think about it, has a distinctive effect on moral judgment.

1. Experiment 1: Moral reflection

People can override intuitive responses. This is illustrated by the ''Cognitive Reflection Test'' (CRT), which consists of questions that elicit intuitively appealing but provably incorrect answers (Frederick, 2005). For example:

> A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

Nearly everyone's first thought is that the ball costs $0.10, but people who consider the problem more thoughtfully discover that the correct answer is $0.05. The CRT was designed as an individual difference measure. Pinillos, Smith, Nair, Marchetto, and Mun (2011) used the CRT to induce more reflective attributions of intentional action. Here we use the CRT in an attempt to induce more reflective moral judgments.

1.1. Participants, methods, and hypotheses

Subjects were recruited, consented, tested, debriefed, and compensated using Amazon's Mechanical Turk[1] (98 females, 52 males; mean age = 34.53, *SD* = 12.92). Subjects completed the three-item CRT and also responded to three ''high-conflict'' (Koenigs et al., 2007) moral dilemmas developed by Greene et al. (2001) in which killing one person will save the lives of several others. One such dilemma follows (see Appendix for the full text of the remaining two dilemmas):

> John is the captain of a military submarine traveling underneath a large iceberg. An onboard explosion has caused the vessel to lose most of its oxygen supply and has injured a crewman who is quickly losing blood. The injured crewman is going to die from his wounds no matter what happens.
>
> The remaining oxygen is not sufficient for the entire crew to make it to the surface. The only way to save the other crew members is for John to shoot dead the injured crewman so that there will be just enough oxygen for the rest of the crew to survive.

Subjects were randomly assigned to complete the CRT either before (CRT-First condition) or after (Dilemmas-First condition) responding to the dilemmas. Subjects evaluated the moral acceptability of the utilitarian action with a binary response (YES/NO), followed by a rating scale (1 = Completely Unacceptable, 7 = Completely Acceptable). No time limits were imposed on responses. Subjects completed the CRT questions and read and

responded to the dilemmas at their own pace. Subjects subsequently completed a brief set of demographic questions.

Success on the CRT requires that one question ultimately override a prepotent intuitive response. According to Greene et al.'s dual-process theory, utilitarian responses to high-conflict dilemmas require a similar process, in this case overriding a countervailing intuitive emotional response. Responding successfully to CRT items reinforces the value of reflection, reminding subjects that their intuitive responses are sometimes incorrect. Thus, we predicted that subjects in the CRT-First condition would become more reflective and consequently judge emotionally aversive utilitarian actions (ones that promote the greater good) to be more acceptable. We also predicted that higher CRT scores would correlate positively with utilitarian judgments. We were agnostic as to whether the latter effect would be due to an induced reflective state, a reflective trait, or both.

1.2. Results

Here, we present data from the rating scales, the more sensitive of our two measures. The three CRT items were scored 0 for incorrect and 1 for correct. Item scores were summed to yield an individual's CRT Score (0–3).

We first examined the effect of completing the CRT before (vs. after) moral judgment by collapsing across the three moral dilemmas, as these responses had good reliability across dilemmas (Cronbach's $\alpha = .71$). Averaging moral judgments in this way yielded a composite moral acceptability rating for each subject. This analysis included only subjects who answered at least one of the three CRT question correctly (92 of 150) because only these subjects show evidence of having reflected on the CRT. (See Appendix for results from the remaining subjects.) Critically, the proportion of subjects in each condition was statistically indistinguishable before and after exclusion (Pre-Exclusion CRT-First: 65 of 150 [43%], Post-Exclusion CRT-First: 41 of 92 [45%], $\chi^2 = .003$, $p = .96$). Results confirmed our first prediction: Subjects in the CRT-First condition judged the utilitarian actions to be more acceptable (CRT-First: $M = 3.77$; Dilemmas-First: $M = 3.25$; $t(90) = 2.03$, $p = .05$, $d = 0.43$).

Again collapsing across all three dilemmas, we examined data from subjects in the CRT-First condition and observed a robust positive correlation between CRT scores and moral acceptability ratings, consistent with our second prediction ($r = .39$, $p = .001$; see Fig. 1). However, there was no such correlation among subjects in the Dilemmas-First condition ($r = -.03$, $p = .8$), suggesting that the aforementioned positive correlation was induced by the CRT and not caused by variability in trait reflectiveness. A Fischer r–z test confirmed that these correlations differed significantly ($z = 2.6$, $p = .01$).

Because our control condition did not include a task prior to the dilemmas task, one might suppose that the reported effects of the CRT are simply effects of performing a task of some non-specific kind. This hypothesis is ruled out by the robust positive correlation observed *within* the CRT-First condition (see Fig. 1). As noted above, this correlation cannot be explained by stable individual differences because there was no such correlation observed in the Dilemmas-First condition. One might then suppose that, in the

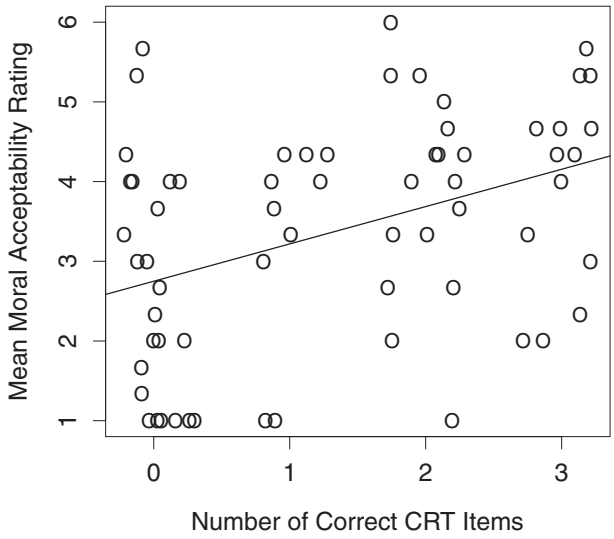

Fig. 1. Scatter plot of Cognitive Reflection Test (CRT) Scores by Moral Acceptability Ratings for the CRT-First condition in Experiment 1. Higher ratings are more utilitarian. A small amount of random jitter has been added to the CRT Scores to aid in visualization.

Dilemmas-First condition, responding to the dilemmas influenced subsequent CRT performance so as to obliterate a latent correlation between CRT scores and moral acceptability ratings. To test this hypothesis, we compared the CRT scores across the two conditions and found no significant effect (CRT-First: $M = 1.32$; Dilemmas-First: $M = 1.16$; $t(148) = 0.83$, $p = .41$).

1.3. Follow-up Study 1

One might hypothesize that the CRT influenced moral judgment, not by inducing reflectiveness, but by inducing positive affect (Valdesolo & DeSteno, 2006) upon solving tricky math problems. To test this hypothesis, we conducted a follow-up study in which subjects were randomly assigned to complete the CRT before or after completing the Positive Affect Negative Affect Schedule (PANAS, Watson, Clark, & Tellegen, 1988).

As in Experiment 1, subjects were recruited using Amazon's Mechanical Turk (46 females, 30 males; mean age = 31.16, $SD = 9.64$). The PANAS was scored in the manner prescribed by Watson et al.—that is, positive and negative affect ratings were averaged within category to yield Positive and Negative Affect Scores. Among subjects who answered at least one CRT question correctly (53 of 76), there was no significant effect of completing the CRT on either positive affect (CRT-First: $M = 2.78$; PANAS-First: $M = 2.84$; $t(51) = -0.29$, $p = .77$) or negative affect (CRT-First: $M = 1.51$; PANAS-First: $M = 1.54$; $t(51) = -0.14$, $p = .89$). The proportion of subjects in each condition was again statistically indistinguishable before and after exclusion (Pre-Exclusion CRT-First: 39 of 76 (51%), Post-Exclusion CRT-First: 28 of 53 (53%), $\chi^2 < .001$, $p = .99$).

1.4. Follow-up Study 2

As noted above, we conducted our primary analyses using composite moral acceptability ratings. We then conducted follow-up analyses of the three individual dilemmas that contributed to the composite ratings: the ''submarine,'' ''crying baby,'' and ''footbridge'' dilemmas (Greene et al., 2001, 2004, 2008). (See above and Appendix for the full text of the dilemmas.) Among subjects who answered at least one CRT item correctly, the effect of completing the CRT before (vs. after) the dilemmas was significant for the submarine dilemma (CRT-First: $M = 5.05$; Dilemmas-First: $M = 4.39$; $t(90) = 1.95$, $p = .05$, $d = 0.41$) and the crying baby dilemma (CRT-First: $M = 3.76$; Dilemmas-First: $M = 2.96$; $t(90) = 2.1$, $p = .04$, $d = 0.44$), but not for the footbridge dilemma (CRT-First: $M = 2.51$; Dilemmas-First: $M = 2.41$; $t(90) = 0.36$, $p = .72$). Likewise, across all subjects in the CRT-First condition we observed robust positive correlations between CRT scores and moral acceptability ratings for the submarine dilemma ($r = .38$, $p = .002$) and the crying baby dilemma ($r = .39$, $p = .001$), but not for the footbridge dilemma ($r = .11$, $p = .4$). For subjects in the Dilemmas-First condition, all three correlations were non-significant: submarine ($r = -.04$, $p = .69$), crying baby ($r = .07$, $p = .51$), footbridge ($r = -.11$, $p = .3$). The difference between correlations (CRT-First vs. Dilemmas-First) was significant for submarine ($z = 2.88$, $p = .004$) and crying baby ($z = 1.95$, $p = .05$), but not for footbridge ($z = 1.32$, $p = .19$).

Thus, these item analyses indicate that the submarine and crying baby dilemmas drove the composite effects reported above, with the footbridge dilemma contributing only minimally. In our original experiment, the mean moral acceptability rating for footbridge ($M = 2.45$) was significantly lower than that of both submarine ($M = 4.41$, $t(149) = -13.26$, $p < .001$, $d = 2.45$) and crying baby ($M = 3.13$, $t(149) = -4.46$, $p < .001$, $d = 0.73$). This suggested a possible floor effect with respect to the CRT manipulation. With this in mind, we conducted a follow-up study, using a version of the footbridge dilemma similar to one used previously by Nichols and Mallon (2006). This dilemma was modified so as to make the utilitarian action more beneficial, and thus bring this dilemma's mean moral acceptability rating closer to the mid-point of the scale and closer to the mean ratings of the other two dilemmas. The revised footbridge dilemma is as follows:

> Half a million people live in a city at the southern end of a valley. At the northern end of the valley is a large dam. Behind the dam is a large lake, several miles wide. A set of train tracks runs across the top of the dam. A container of explosives has accidentally been left on the tracks.
>
> A runaway trolley is speeding down these tracks. If nothing is done, the trolley will soon collide with the explosives, creating an explosion that will cause the dam to burst. The water from the lake will flood the valley, including the city below. Thousands of people will die as a result.
>
> It is possible to avoid these deaths. There is a footbridge above the tracks in between the runaway trolley and the dam. On this footbridge is a railway workman wearing a large,

heavy backpack. Joe is a bystander who understands what is going on and who happens to be standing behind the workman on the footbridge. Joe sees that the only way to prevent the dam from bursting is to push the workman off of the footbridge and onto the tracks below. If he does this, the trolley will collide with the workman, and the combined weight of the workman and the backpack will be enough to stop the trolley. This will prevent the deaths of thousands of people in the city. However, the collision will cause the death of the workman with the backpack.

Note that Joe cannot stop the trolley by jumping onto the tracks himself because he is not heavy enough to stop the trolley. Nor is there enough time for him to remove the workman's backpack and put it on himself.

Subjects (75 females, 42 males, 1 gender unspecified; mean age = 32.93, SD = 14.98) were recruited through TestMyBrain.org and participated on a volunteer basis. Otherwise, all methods were identical to those employed in the original experiment. As expected, the mean moral acceptability rating was close to the midpoint of the scale (M = 3.53). Among participants who answered at least one CRT item correctly (72 of 118), the effect of completing the CRT before (vs. after) the dilemmas was non-significant (CRT-First: M = 4.03; Dilemmas-First: M = 3.77; $t(70) = 0.58$, $p = .56$). As before, the proportion of subjects in each condition was statistically indistinguishable before and after exclusion (Pre-Exclusion CRT-First: 61 of 118 (52%), Post-Exclusion CRT-First: 37 of 72 (51%), $\chi^2 = .01$, $p = .91$). Among all subjects in the CRT-First condition, we observed the predicted positive correlation between CRT scores and moral acceptability ratings ($r = .25$, $p = .05$). However, here we also observed a marginally significant correlation between CRT scores and moral acceptability ratings among all subjects in the Dilemmas-First condition ($r = .25$, $p = .06$). The difference between the correlations was non-significant ($z = 0.04$, $p = .97$). Pooling the results from both conditions naturally yielded a correlation of comparable effect size ($r = .25$) and higher significance ($p = .007$).

Thus, while the overall results of Experiment 1 indicate that it is possible to increase utilitarian judgment by inducing a more reflective state of mind, these additional results indicate that variability in trait reflectiveness (a non-induced behavioral tendency to be more or less reflective) is also associated with greater utilitarian judgment.[2] All of the above results are consistent in associating utilitarian judgments with greater reflectiveness.

2. Experiment 2: Reasoned reflection

Experiment 1 indicates that reflection can influence moral judgment when people are induced to distrust their immediate intuitive responses. However, this experiment does not specifically address the role of moral reasoning.

Two hallmarks of reasoned reflection are (a) sensitivity to argument strength, and (b) extended temporal duration. To obtain strong evidence for reasoned reflection, both hallmarks must be observed. A judgment process that fails to distinguish between strong and

weak arguments could hardly be called ''reasoning,'' no matter how long it takes. Likewise, a judgment process that is sensitive to argument strength, but that happens immediately, may be characterized as intuitive. In Experiment 2, we therefore examined the effects of both argument strength and deliberation time on moral judgment.

2.1. Participants, methods, and hypothesis

Subjects (79 females, 61 males, 2 gender unspecified; mean age = 23.69, *SD* = 7.46, 1 age unspecified) were recruited from the Harvard Psychology Department Study Pool.

Subjects read on a computer a vignette describing an episode of consensual incest between a brother and sister (Haidt, Bjorklund, & Murphy, unpublished data):

> Julie and Mark are sister and brother. They are traveling together in France one summer vacation from university. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other.

The vignette is designed to minimize harm-based (utilitarian) reasons for condemning the couple. We chose this vignette because it is known for eliciting emotionally driven condemnation that resists reasoned persuasion (Haidt, 2001).

After reading the vignette at their own pace, subjects were presented with either a strong or a weak argument defending the counterintuitive claim that incest is morally acceptable in this and similar cases:

> *Strong Argument*: For most of our evolutionary history, there were no effective contraceptives, and so if siblings slept together they might conceive a child. Children born of such closely related parents would have a lower than normal likelihood of surviving. Thus, feelings of disgust toward incest probably evolved to prevent such children from being born. But in Julie and Mark's case, two kinds of contraception were used, so there was no chance of conceiving a child. The evolutionary reason for the feeling of disgust is, therefore, not present in Julie and Mark's case. Any disgust that one feels in response to Julie and Mark's case cannot be sufficient justification for judging their behavior to be morally wrong.

> *Weak Argument*: A brother–sister relationship is, by it's nature, a loving relationship. And making love is the ultimate expression of love. Therefore, it makes perfect sense for a brother and sister, like Julie and Mark, to make love. If more brothers and sisters were to make love, there would be more love in the world, and that is a good thing. If brothers and sisters were not supposed to make love, then they wouldn't be sexually compatible, and yet they are. Brothers and sisters who don't want to make love should at least try it

once. There is nothing wrong with trying something once. Thus, it wasn't morally wrong for Julie and Mark to make love.

Subjects read the argument at their own pace. Approximately half of the subjects were then randomly assigned to think about the argument for an additional 2 min, during which time the argument remained on the screen. Following the argument, all subjects rated the moral acceptability of Julie and Mark's behavior using a 1–7 scale. The experiment thus employed a 2 × 2, between-subjects design, which randomly varied both the strength of the argument presented (strong vs. weak) and the minimum deliberation time (immediate response allowed vs. required 2-min delay). Subjects also answered a small number of personality questionnaires (see Appendix) along with a brief set of demographic questions, and were provided with a debriefing form.

Our hypothesis was that reasoned reflection would influence moral judgment. Specifically, we predicted an interaction such that the effect of argument strength would be stronger when increased deliberation time encouraged subjects to reflect on the arguments.

2.2. Results

We analyzed the moral acceptability ratings using a two-way, between-subjects ANOVA. This revealed a main effect of argument strength ($F(1, 136) = 9.09$, $p = .003$, $\eta^2_{\text{partial}} = .06$), consistent with our designation of the strong and weak arguments as such. In addition, we observed a marginally significant main effect of deliberation time ($F(1, 136) = 2.77$, $p = .1$, $\eta^2_{\text{partial}} = .02$). Importantly, we observed the predicted interaction between argument strength and deliberation time ($F(1, 136) = 4.95$, $p = .03$, $\eta^2_{\text{partial}} = .04$; see Fig. 2). Planned contrasts revealed a predicted simple effect of argument strength within the delayed response condition ($F(1, 136) = 9.07$), $p = .003$, $\eta^2_{\text{partial}} = .06$) and no effect of argument strength in the immediate response condition ($F(1, 136) = 0.15$, $p = .85$). Thus, our hypothesis was confirmed in a surprisingly strong form. Reflection not only increased the effect of argument strength. There was no effect of argument strength when reflection was not encouraged.

3. Discussion

We examined the roles of reflection and reasoning in moral judgment. Experiment 1 documented the influence of reflection on moral judgment by inducing people to be more reflective. In general, this reflectiveness manipulation increased utilitarian moral judgment, although there was one item for which this effect was not reliably observed. A follow-up study indicated that trait reflectiveness is also associated with increased utilitarian judgment. Both results are consistent with Greene et al.'s (2001, 2004, 2008) dual-process theory of moral judgment. However, it is unclear why some dilemmas yielded effects of reflective states, but not reflective traits, while at least one dilemma exhibited the reverse pattern. We leave this as a matter for future research.

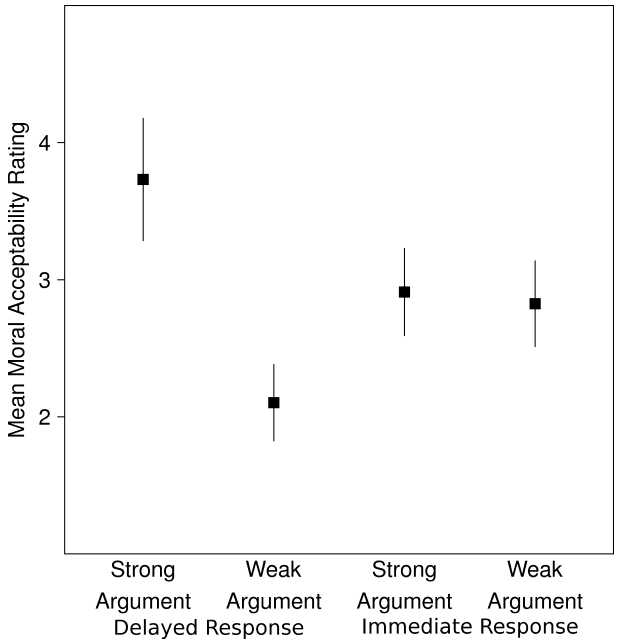

Fig. 2. Mean Moral Acceptability Ratings by condition for Experiment 2. Higher ratings are more utilitarian. Error bars represent standard error of the mean. Interaction: $p = .03$.

We have hypothesized that these utilitarian judgments, in addition to involving moral reflection, also involve moral reasoning, that is, the deliberate application of a utilitarian moral principle favoring the action that saves the most lives. Experiment 1, however, did not specifically examine reasoning. Experiment 2 examined both reflection and reasoning by examining the effects of argument strength and deliberation time on moral judgment. Consistent with the influence of reasoned reflection, we found that a strong argument was more persuasive than a weak one, but only when subjects were encouraged to reflect. These results are consistent with a recent study (Suter & Hertwig, 2011) showing that decreased deliberation decreases utilitarian judgment. The present results demonstrate a parallel effect in which increased deliberation influences moral judgment, making judgments more consistent with utilitarian principles. This effect depends critically on argument strength, thus implicating moral reasoning. Based on our reading of the literature (Paxton & Greene, 2010), these results provide the strongest evidence to date for the influence of reflection and reasoning on moral judgment.

As noted above, understanding the respective roles of emotion, reflection, and reasoning in moral judgment has been a central issue in contemporary moral psychology. Haidt's SIM emphasizes the role of emotional intuition, while deemphasizing (although not dismissing) the roles of reflection and reasoning. Greene et al.'s dual-process theory gives moral reasoning—especially utilitarian reasoning—more prominent billing. One may interpret the present results as supporting previously unsupported (or under-supported) components of

the SIM. Alternatively, one may interpret our results as favoring the dual-process theory over the SIM. To some extent, this interpretive question is a matter of imprecise evidential bookkeeping. Two points, however, deserve attention, both of which bear on broader questions concerning the possibility and nature of moral progress (Bloom, 2010; Greene, 2012).

First, Experiment 1 indicates that moral reflection and reasoning are not simply matters of managing competing intuitions. There is no reason to think that the CRT manipulation influenced moral judgment by favoring one intuition over another intuition. Rather, the CRT manipulation was designed to induce a general distrust of intuition. Thus, its efficacy implies that the induced utilitarian judgments are in some sense counter-intuitive and not simply driven by competing intuitions. We note that this effect of induced counter-intuitive judgment was observed not in trained moral philosophers, or even in college students, but in a diverse sample of Internet users. Interestingly, this effect required that these individuals understand, on some level, which of their judgments are intuitive, implying a kind of metacognitive knowledge. Second, Experiment 2 demonstrated the persuasive power of an abstract argument based on an evolutionary theory of the origins of the incest taboo. Assuming that this argument did little to reduce the emotional aversiveness of incest, these results suggest that it is possible to persuade people by appealing to their ''heads'' as well as their ''hearts.''

Notes

1. According to Buhrmester, Kwang, and Gosling (2011), subjects recruited through Mechanical Turk are more representative than the average Internet sample and yield data at least as reliable as those obtained in the lab.
2. In an unpublished manuscript, Hardman reports a similar correlation when randomly intermingling CRT items with moral dilemmas.

Acknowledgments

We thank Laura Germine (TestMyBrain.org) and Eva Liou for assistance in data collection, and Joshua Knobe for advice on experimental design. We thank Jonathan Haidt and an anonymous reviewer for helpful comments. This work was supported by a National Science Foundation Graduate Research Fellowship awarded to J.M.P. and NSF SES-082197 8 awarded to J.D.G.

References

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition*, *108*(2), 381–417.

Bloom, P. (2010). How do morals change? *Nature*, *464*(7288), 490.

Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data. *Perspectives on Psychological Science*, *6*(1), 3–5.

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.

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

Greene, J. D. (2012). *The moral brain and how to use it*. New York: Penguin.

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., Nystrom, L., Engell, A., Darley, J., & Cohen, J. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(2), 389–400.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (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*(4), 814–834.

Haidt, J. (2003). The emotional dog does learn new tricks: A reply to Pizarro and Bloom (2003). *Psychological Review*, *110*(1), 197–198.

Haidt, J., & Kesebir, S. (2010). Morality. In S. Fiske, D. Gilbert, & G. Lindzey (Eds), *Handbook of social psychology* (5th ed., pp. 797–832). Hoboken, NJ: Wiley.

Inbar, Y., Pizarro, D., Knobe, J., & Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion*, *9*(3), 435–439.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Kohlberg, L. (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 347–480). Chicago, IL: Rand McNally.

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.

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

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, *109*(2), 134–141.

Paxton, J. M., & Greene, J. D. (2010). Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, *2*(3), 511–527.

Petty, R., & Cacioppo, J. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 123–205.

Pinillos, N., Smith, N., Nair, G., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. *Mind & Language*, *26*(1), 115–139.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review*, *110*(1), 193–196.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653–660.

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science*, *19*(12), 1219–1222.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, UK: Cambridge University Press.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science*, *17*(6), 476–477.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science*, *18*(8), 689.

Watson, D., Clark, L., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784.

Appendix

This document contains testing materials and additional analyses from both Experiments 1 and 2 described in our article ''Reflection and Reasoning in Moral Judgment.''

1. Experiment 1

1.1. Moral dilemmas

The three moral dilemmas used in this experiment were versions of the *Submarine*, *Crying Baby*, and *Footbridge* dilemmas developed by Greene et al. (2001). Subjects did not see these labels, as they are used here only to identify the dilemmas. The first question associated with each dilemma was given a binary response (YES/NO), while the second was made using a Likert-type rating scale (1 = Completely Unacceptable, 7 = Completely Acceptable). The full text of the submarine dilemma can be found in the main text. The full text of the remaining two dilemmas follows:

Crying Baby

Enemy soldiers have taken over Jane's village. They have orders to kill all remaining civilians. Jane and some of her townspeople have sought refuge in the cellar of a large house. Outside they hear the voices of soldiers who have come to search the house for valuables.

Jane's baby begins to cry loudly. She covers his mouth to block the sound. If she removes her hand from his mouth, his crying will summon the attention of the soldiers, who will kill her, her child, and the others hiding out in the cellar. To save herself and the others, she must smother her child to death.

Footbridge

A runaway trolley is heading down the tracks toward five railway workmen, who will be killed if the trolley proceeds on its present course. Jane is on a footbridge over the tracks, in between the approaching trolley and the five workmen. Next to her on this footbridge is a lone railway workman, who happens to be wearing a large, heavy backpack.

The only way to save the lives of the five workmen is for Jane to push the lone workman off the bridge and onto the tracks below, where he and his large backpack will stop the trolley. The lone workman will die if Jane does this, but the five workmen will be saved.

1.2. Excluded subjects

The first analysis under Experiment 1 included only subjects who showed evidence of having reflected on the Cognitive Reflection Task (CRT) by answering at least one CRT question correctly (92 of 150), and found that such subjects in the CRT-First condition were more utilitarian than those in the Dilemmas-First condition, consistent with the dual-process theory of moral judgment (Greene et al., 2001; Greene, 2009). In addition, we observed a marginally significant trend in the opposite direction for subjects who failed to answer at least one CRT question correctly, and thus showed no evidence of having reflected on the CRT (CRT-First: $M = 2.68$; Dilemmas-First: $M = 3.37$; $t(56) = -1.74$, $p = .09$). This trend is likewise consistent with the dual-process theory, which associates controlled cognitive processing with utilitarian moral judgment, and automatic intuitive processing with non-utilitarian moral judgment. More specifically, it appears that completing a set of apparently easy math problems emboldened these subjects to rely more heavily on their automatic intuitions, resulting in less utilitarian judgment.

2. Experiment 2

2.1. Personality scales

Subjects completed three personality scales in the following order prior to completing the experiment: the Cognitive Reflection Test (Frederick, 2005), the Disgust Scale – Revised (Haidt et al., 1994; Olatunji et al., 2007), and the Rational-Experiential Inventory (Epstein et al., 1996). Analyses of the questionnaire data were inconclusive and therefore the results are not reported here.

References

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *62*(2), 390–405.

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

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*(5537), 2105–2108.

Greene, J.D. (2009). The cognitive neuroscience of moral judgment. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences 4th edition*, pp. (987–1002). Cambridge, MA: The MIT Press.

Haidt, J., McCauley, C., & Rozin, P. (1994). Individual differences in sensitivity to disgust: A scale sampling seven domains of disgust elicitors. *Personality and Individual Differences*, *16*(5), 701–713.

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

Olatunji, B. O., Williams, N. L., Tolin, D. F., Abramowitz, J. S., Sawchuk, C. N., Lohr, J. M., & Elwood, L. S. (2007). The disgust scale: Item analysis, factor structure, and suggestions for refinement. *Psychological Assessment*, *19*(3), 281–297.