Topics in Cognitive Science (2010) 1–17

ISSN: 1756-8757 print / 1756-8765 online
DOI: 10.1111/j.1756-8765.2010.01096.x

Moral Reasoning: Hints and Allegations

Joseph M. Paxton, Joshua D. Greene

Department of Psychology, Harvard University

Received 17 June 2009; received in revised form 05 November 2009; accepted 27 January 2010

Abstract

Recent research in moral psychology highlights the role of emotion and intuition in moral judgment. In the wake of these findings, the role and significance of moral reasoning remain uncertain. In this article, we distinguish among different kinds of moral reasoning and review evidence suggesting that at least some kinds of moral reasoning play significant roles in moral judgment, including roles in abandoning moral intuitions in the absence of justifying reasons, applying both deontological and utilitarian moral principles, and counteracting automatic tendencies toward bias that would otherwise dominate behavior. We argue that little is known about the psychology of moral reasoning and that it may yet prove to be a potent social force.

Keywords: Dual-process model; Moral judgment; Moral reasoning; Social intuitionist model

1. Introduction

The following is based on a true story:

> Greg and Adam are high school buddies. Adam is a vegetarian. Greg is not. Both enjoy eating meat, but Adam has given it up after concluding that eating meat is morally wrong. Over many months, Adam and Greg argue about the ethics of eating meat. Adam agrees with Greg that hamburgers taste better than veggie burgers, but he argues that the additional enjoyment that we humans derive from eating meat is not enough to justify the suffering and ultimate death inflicted on animals such as cows. Greg is not easily convinced. He observes that eating meat is perfectly natural, pointing to his canine teeth. Adam replies that many things, such as wars of aggression, may be perfectly natural, but that such things are not necessarily right. Greg points out that the animals he eats owe

Correspondence should be sent to Joseph M. Paxton, 33 Kirkland St., Room 1484, Department of Psychology, Harvard University, Cambridge, MA 02138. E-mail: jpaxton@wjh.harvard.edu

> their very existence to the demands of consumers such as himself. Adam replies that most animals raised for food live miserable lives and would be better off not existing. Through the course of many such discussions, Greg's mind is changed and he, too, becomes a vegetarian.

What is going on in this kind of exchange? For decades, the field of moral psychology emphasized the role of reasoning in moral judgment (Kohlberg, 1969; Smetana & Killen, 2006; Turiel, 1983), while more recent research (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001; Haidt, Koller, & Dias, 1993) emphasizes the power and prevalence of emotionally based moral intuition. There is a sense in which the above exchange must be a case of moral reasoning: Adam gave Greg *reasons* for becoming a vegetarian, and Greg changed his stated beliefs and behavior in response to those reasons. But there are two important ways in which the role of moral reasoning in this exchange remains uncertain. First, did Adam arrive at his current, pro-vegetarian stance because he was himself compelled by the arguments he gave, or are his arguments mere rationalizations for his preformed judgment? Second, independent of how Adam's attitudes were formed, did Adam change Greg's mind by modifying Greg's *intuitions* about eating meat? Or did Adam change Greg's mind by providing Greg with a reasoned argument that exerted an influence independent of, or even in spite of, Greg's intuitions? In more colloquial terms, did Adam change Greg's mind by appealing to his ''heart'' or his ''head?'' Is there a legitimate distinction between these two types of persuasion, and, if there is, what evidence is there concerning the reality, prevalence, and significance of each type of persuasion? In what follows we address these questions.

2. Two theories of moral judgment

We begin with Haidt's (2001) highly influential framework for understanding moral psychology: the social intuitionist model (SIM). The SIM consists of a set of causal ''links'' connecting three types of psychological process: intuition, judgment, and reasoning (Fig. 1). The backbone of the SIM consists of two links. The ''intuitive judgment'' link posits that one's judgments are driven primarily by one's intuitions, while the ''post-hoc reasoning'' link posits (contrary to traditional rationalist models) that one's reasoning is driven primarily by one's judgment, rather than the other way around. These two links are supplemented by weaker links that allow reasoning to, on occasion, exert a causal influence on judgment. The ''reasoned judgment'' link allows one's reasoning to directly influence one's judgment, while the ''private reflection'' link allows one's reasoning to influence one's judgment by modifying one's intuitions. The ''social'' in the SIM comes from two additional links: the ''reasoned persuasion'' link, by which one person's reasoning influences another's judgment by influencing that person's intuition, and the ''social persuasion'' link, by which one's judgment, in the absence of explicit attempts at reasoning, influences another's judgment by modifying that person's intuition.

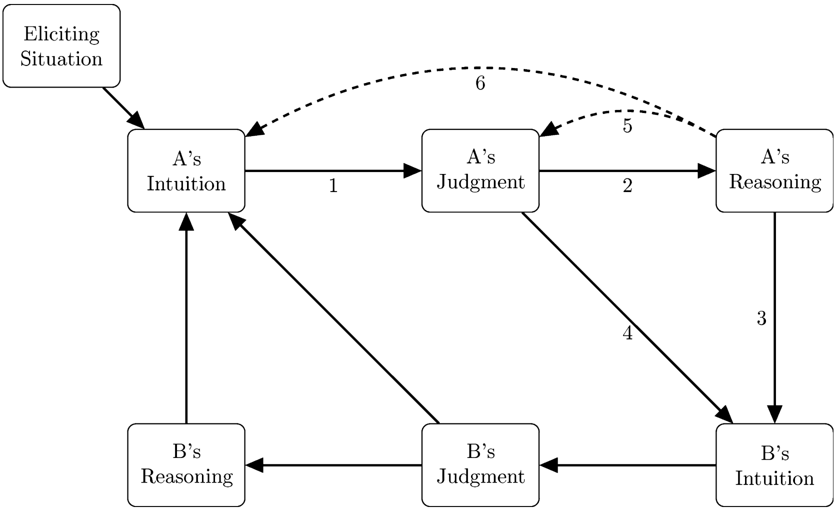

Fig. 1. Haidt's (2001) social intuitionist model (SIM) consists of six links describing causal connections among moral intuitions, moral judgments, and episodes of moral reasoning: (1) intuitive judgment, (2) post-hoc reasoning, (3) reasoned persuasion, (4) social persuasion, (5) reasoned judgment, and (6) private reflection. Dashed lines indicate links that are rarely used.

Greene and colleagues (Greene, 2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001) have developed an alternative dual-process model of moral judgment that is in many ways consistent with the SIM. Greene's model posits two natural, ubiquitous, and qualitatively different modes of moral thinking that depend on dissociable, and in some cases competing, systems in the brain (Fig. 2). According to Greene, deontological moral judgments, judgments that are naturally regarded as reflecting concerns for rights and duties, are driven primarily by intuitive emotional responses. At the same time, Greene et al. argue that utilitarian⁄consequentialist judgments, judgments aimed at promoting the greater good, are supported by controlled cognitive processes that look more like moral reasoning. For example, when people are confronted with the possibility of saving five people by pushing one person in front of a runaway trolley, it appears that the inclination to respect the ''rights'' of the would-be victim is driven by emotional responses that depend on the ventromedial prefrontal cortex (Ciaramelli, Muccioli, Ladavas, & di Pellegrino, 2007; Greene et al., 2001, 2004; Koenigs et al., 2007; Mendez, Anderson, & Shapira, 2005), while the countervailing utilitarian judgment is driven by controlled cognitive processes that depend on the dorsolateral prefrontal cortex (Greene et al., 2001, 2004, 2008).

For present purposes, there are two critical differences between Haidt's SIM and Greene's dual-process model. First, while the SIM posits that reasoned judgment within an individual is, ''rare, occurring primarily in cases in which the intuition is weak and processing capacity is high,'' Greene's dual-process model allows that moral reasoning—especially utilitarian⁄consequentialist reasoning—may be a ubiquitous feature of moral common sense. Second, according to the SIM, social influence on moral judgment only occurs when

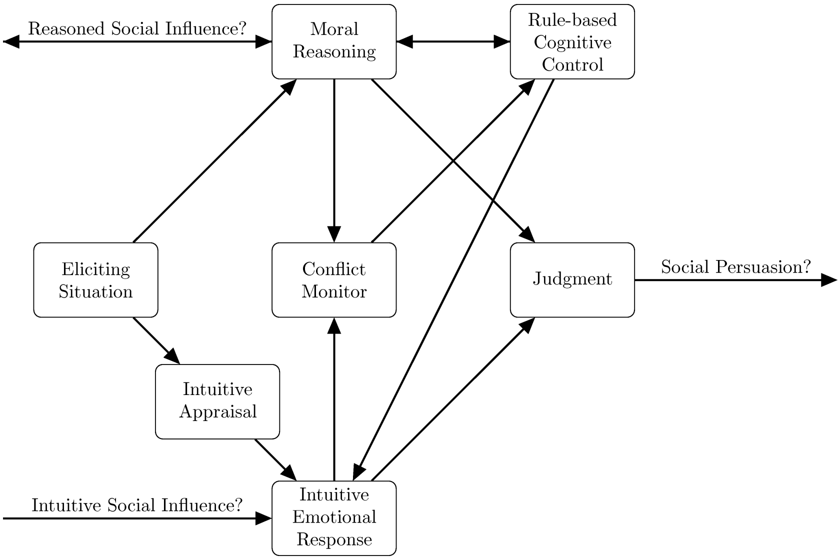
Reasoned Social Influence?
Moral Reasoning
Rule-based Cognitive Control
Eliciting Situation
Conflict Monitor
Judgment
Social Persuasion?
Intuitive Appraisal
Intuitive Social Influence?
Intuitive Emotional Response

Fig. 2. According to Greene et al.'s (2001, 2004, 2008) dual-process model, moral judgments are driven by both intuitive emotional responses and controlled cognitive responses. This model differs from the SIM in two critical ways. First, it emphasizes the role of rule-based, controlled cognitive processes, especially the conscious application of utilitarian moral principles. Second, it allows that social influence may occur when people directly engage one another's capacities for moral reasoning, that is, the conscious evaluation of moral judgments/behaviors for their consistency with moral principles and other moral commitments.

one person succeeds in modifying another's intuition. In other words, the SIM includes no social counterpart to the ‘‘reasoned judgment'' link, which would allow one person's moral reasoning to influence another's moral judgment directly, without first modifying the target's intuition. This last difference, while perhaps seeming trivial, is actually of great social significance, with important implications for the practice of democracy and the pursuit of social progress. If the SIM is correct, then attempting to engage others through their ability to reason, aiming messages at the ‘‘head'' rather than the ‘‘heart,'' is an exercise in futility. Indeed, depending on what one means by ‘‘reasoning,'' one might say that the SIM does not really allow for ‘‘reasoned persuasion'' at all. According to the SIM, we can say things in hopes of modifying one another's moral intuitions, but it is impossible to convince someone to put her moral intuitions aside and, instead, reach an alternative, counter-intuitive conclusion based on reasoning. According to the SIM, it is impossible for Adam to turn Greg into a vegetarian (or a supporter of healthcare reform, or gay marriage, or anything else) without first changing the way Greg *feels*. In contrast, the dual-process model suggests that Adam may be able to change Greg's mind, and even his ‘‘heart,'' by starting with his ‘‘head,'' by targeting the system for controlled cognition that is based in the dorsolateral prefrontal cortex. Note that the disagreement here is not over the importance—even dominance—of emotion and intuition in moral psychology and moral discourse. The disagreement is over the reality, or even the mere possibility, of what we regard as genuinely reasoned moral discussion, that is, discussion in which one person's capacity for moral reasoning is directly engaged with another's.

Against this theoretical backdrop, this paper has two goals. First, with respect to the existence and prevalence of individual ''reasoned judgment'' our aim is simply to present and consider the available evidence. Here, the SIM and dual-process theories differ only in degree, and the decision to favor one model over the other may ultimately come down to one's preferred accounting system. With respect to ''reasoned persuasion,'' the disagreement is more substantial, but the available evidence is more scant. Here, our aim is to review the available evidence, however limited it may be, and to lay a conceptual foundation for research that could provide more conclusive evidence.

3. What is moral reasoning?

Haidt defines moral reasoning as ''conscious mental activity that consists of transforming given information about people in order to reach a moral judgment'' (Haidt, 2001). While this definition provides a useful starting point, we believe that it may be too broad, as it would allow any conscious thought process (about people) that affects moral judgment to count as ''moral reasoning.''

For example, we might engage in conscious reasoning to determine whether Oswald shot JFK, and such reasoning may affect our ultimate judgment concerning Oswald's moral guilt or innocence (Bucciarelli, Khemlani, & Johnson-Laird, 2008). There's little doubt that we often need to make factual inferences in order to reach a moral judgment, that these inferences are instances of reasoning, and that they have an important influence on moral judgment. But this is not moral reasoning in the sense that has been, and remains, controversial in moral psychology.

A related problem concerns the classification of mental activity that is specifically moral, but that may not seem sufficiently reasoned to qualify as ''reasoning.'' If, for example, one consciously thinks to oneself, or says to another in an attempt at persuasion, ''Anti-war protesters are communist, fascist, pigs who should go back to Russia!'' that may qualify as ''conscious mental activity that consists of transforming given information about people in order to reach a moral judgment,'' but many of us would not want to count this as ''moral reasoning.'' And the same goes for more congenial moral communications, such as many of those delivered in Martin Luther King Jr.'s famous ''I Have a Dream'' speech, in which persuasion is effected primarily (though not exclusively) through metaphor and imagery. Haidt (2001) regards such communication as ''reasoned persuasion.'' We think this classification is questionable, but our task here is not to quibble over definitions. The more substantive point is that there is, or may be, a kind of moral persuasion that is more reasoned than this—more ''head'' and less ''heart.''

Take for example, Adam's argument with Greg. Adam might have said, ''Greg, I have a dream… that one day cows and humans will stroll together through the pastures of peace…'' But Adam takes a different approach. Adam points out that animals raised for meat suffer, but he goes further. He claims that the suffering animals experience is not outweighed by the enjoyment that humans gain by eating meat rather than vegetable products, explicitly invoking a utilitarian/consequentialist principle. Greg does not dispute Adam's utilitarian calculus, but attempts to render it irrelevant by claiming that eating meat is natural, implicitly appealing to the principle that what is natural is right or good. And Adam replies by observing that things that Greg himself regards as bad, such as wars of aggression, may be natural, thus invalidating the principle to which Greg has implicitly appealed. And so on.

What, then, is the difference between ''I have a dream…'' and this kind of argumentative persuasion? One might suggest that, ultimately, there is no difference, that Adam's arguments are nothing more than attempts to push intuitively compelling metaphors and images into Greg's head. That may be the case, but, alternatively, it may be that Adam is changing Greg's mind by appealing to his capacity to reason, and not (primarily) by modifying Greg's intuitions. If that's the case, what exactly does it mean for one person to engage another's reasoning?

We propose that moral reasoning, in this more restricted sense, involves an attempt to compel another individual (or oneself) to accept a moral conclusion *on pain of inconsistency*. For the sake of clarity, we will refer to this more restricted kind of moral reasoning as ''Moral Reasoning,'' defined as follows:

> Moral Reasoning: Conscious mental activity through which one evaluates a moral judgment for its (in)consistency with other moral commitments, where these commitments are to one or more moral principles and (in some cases) particular moral judgments.[1]

This definition explains why ''I have a dream…'' need not count as Moral Reasoning. One can be moved to pursue King's dream of racial harmony without any conscious recognition that to do otherwise would be inconsistent with one's other moral commitments. In contrast, Adam's persuasive strategy is to point out that eating meat is inconsistent with a utilitarian principle. This is different from saying, ''But think of the poor suffering animals!'' Adam is saying, ''Think of the poor suffering animals, but also think of the happy meat-eating humans. Do the math, and come to the inexorable conclusion that the animal suffering caused by eating meat is greater than the human happiness caused by eating meat. If you accept the moral principle that we should not do things that make the world an overall less happy (and more suffering) place, then you, too, should be opposed to eating meat.'' Such a strategy can only be effective in response to an interlocutor who is capable of recognizing logical inconsistency and motivated to avoid it. ''I have a dream,'' in contrast, can be successful in response to anyone who is capable of sharing another's dream.

Our hypothesis is that Moral Reasoning not only happens, but that, for all we know, it may be a pervasive and important aspect of our moral psychology, even if it is relatively rare compared to more intuitive moral reasoning (Pizarro & Bloom, 2003). In what follows we consider the evidence, limited though it may be, that supports this hypothesis.

4. Inducing moral reasoning

To determine whether people engage in Moral Reasoning, one might begin by examining how people behave when asked to take a more rational approach to moral judgment. Pizarro, Uhlmann, and Bloom (2003) took this approach in a study employing scenarios like this:

> Barbara wants to kill her husband, John. When they are eating at a restaurant, Barbara slips some poison into John's dish while he isn't looking. Unbeknownst to Barbara, the poison isn't strong enough to kill her husband. However, it makes the dish taste so bad that John changes his order. When he receives his new order, it contains a food that John is extremely allergic to, and which kills him within minutes.

Here, Barbara is a cause of John's death. She intends to kill him, and he does indeed die, but John does not die in exactly the way that Barbara intends. Pizarro and colleagues found that, under normal conditions, subjects assigned less blame to the agent in ''causally deviant'' scenarios like this one, relative to casually normal scenarios in which harm unfolds exactly as the agent intended. However, when subjects were instructed to first make a ''rational, objective judgment,'' they discounted blame less than when they were first instructed to go with their ''intuitive, gut feeling.''

Even if one regards the effect observed here as an effect of experimenter demand, the question remains, why did the demand to be more ''rational'' and ''objective'' lead subjects to alter their judgments as they did? They must think there is something irrational about blaming Barbara less because her murderous plot failed to unfold exactly as intended. In other words, the subjects appear to be applying some kind of normative standard, a principle, for evaluating judgments of blameworthiness. Taking a bit of liberty, they may be thinking something like this: ''Barbara seems less blameworthy when she causes John's death in this weird, coincidental way. But does that really make sense? Her intention is just as bad. The result is just as bad. And she is causally responsible for the bad result. The only reason I can see for reducing her blame is that she just seems less blameworthy. But that does not sound like a very good reason. And I was just asked to be rational, so…''

In other words, this appears to be a case of Moral Reasoning, a case in which a judgment is not merely altered through conscious mental activity, but altered in a *principled* way, so as to be consistent with one's other moral commitments (''Deviant causation is not on the list of things that matter for blame''). This interpretation does not necessarily pose a problem for the SIM. One might simply regard this phenomenon as evidence—much needed evidence—that private ''reasoned judgment'' sometimes happens. And it is worth noting that this does not appear to be a case in which reason resolves a dispute between two competing intuitions. Rather, it appears to be a case in which intuition and reason conflict, as suggested by the fact that the instruction to be ''rational'' reliably pushes subjects in a particular direction.

These results challenge the SIM in a more serious way if the effect is viewed as a product of social interaction. The subject hears the experimenter's instruction to be ''rational'' and chooses an answer that he believes will stand up better to social scrutiny. If that's the case, then it appears that one person can influence another person's judgment, not by modifying the target's intuition, but by appealing to the target's capacity to reason, to formulate judgments that are consistent with their other moral commitments.

Of course, it does not follow from this that the target has actually changed his judgment. These subjects may simply be telling the experimenter what they think she wants to hear. However, even if this effect is driven in part by demand, these results still show that one can influence another's judgment simply by urging that person to be ''rational.'' This suggests that people may well have a capacity to engage one another with Moral Reasoning: We know what we're supposed to do when told to be ''rational'' because this is something that we, at least occasionally, do.

5. Application of deontological moral principles

Among the moral principles discussed and applied by professional and amateur philosophers, the most prominent may be deontological principles. Deontological moral principles prohibit or allow certain types of actions based on the features of those actions, as opposed to their consequences. For example, according to the *action principle*, harm caused by an action is less morally acceptable than harm caused by an omission.

A pair of recent studies (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Cushman, Young, & Hauser, 2006) provides evidence that subjects apply deontological moral principles when making moral judgments in response to hypothetical ''trolley'' scenarios (Foot, 1978; Greene et al., 2001; Mikhail, 2010; Petrinovich, O'Neill, & Jorgensen, 1993; Thomson, 1985). In Cushman et al.'s study, matched pairs of scenarios were constructed in which harm is caused either through action (e.g., hitting a lever that turns a runaway trolley away from five people and onto one person) or through omission (e.g., refraining from hitting a lever that will turn a runway trolley away from one person and onto five people). In the first phase of the experiment, subjects morally evaluated a series of harmful actions and omissions as described above. In the second phase, subjects were asked to justify their responses to matched pairs of previously evaluated action/omission scenarios. The experimenters then coded each justification with respect to whether the subject provided a clear appeal to the action principle.

Subjects' judgments in the first phase of the experiment did indeed conform to the action principle, with harmful actions evaluated less favorably than harmful omissions. Moreover, in the second phase of the experiment, subjects provided clear appeals to the action principle for 81% of the scenario pairs. This indicates that subjects could have consciously applied the action principle in generating their judgments, as they were able, in the end, to consciously articulate the action principle.

In an unpublished reanalysis of these data (F. Cushman, personal communications, January 2009), the authors asked whether subjects who successfully articulated the action principle after the first phase of the experiment were more likely to have consistently conformed to the action principle in the first phase. That is indeed what happened, suggesting that at least some subjects were consciously applying the action principle during the first phase of the experiment.

However, it is also possible that subjects whose judgments best conformed to the action principle during the first phase were more likely to articulate the action principle in the second phase, not because they had consciously applied that principle earlier, but simply because their past behavior made the action principle more accessible during the second phase.

An independent functional magnetic resonance imaging (fMRI) study (Borg et al., 2006) provides convergent evidence for the hypothesis that people consciously apply the action principle in making moral judgments. In this study, subjects were scanned while making judgments about moral dilemmas similar to those listed above. For moral dilemmas involving the action principle (as compared to analogous nonmoral scenarios), the dorsolateral prefrontal cortex (DLPFC) was identified as area of maximal activation. While the activation of the DLPFC does not guarantee that the process in question is conscious, the DLPFC is typically implicated in conscious processing (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006). More specifically, the DLPFC is widely implicated in controlled cognitive tasks, such as the color-naming Stroop (1935) task, in which one must apply a conscious decision-rule (typically one provided explicitly by the experimenter) in the face of a countervailing prepotent response (MacDonald, Cohen, Stenger, & Carter, 2000; Miller & Cohen, 2001).

Thus, these fMRI data provide further support for the interpretation offered by Cushman et al. for their justification results. If the post-hoc interpretation were correct, one would not expect to see neural activity associated with Reasoning and controlled cognition engaged until after the decision had already been made. But, in this case, the timing of the activation corresponds to the point of decision, rather than points after the decision. Thus, the activation in question is more likely to be indicative of a Reasoning process that led to the decision than a rationalizing process that came after the decision.

In light of these convergent findings, there is good reason to think that, at least in some cases, people consciously apply deontological principles such as the action principle. Even if this is true, however, one might still regard such Reasoning as a kind of post-hoc rationalization. This is because the principle in question may itself have been formulated merely as an attempt to rationalize prior moral intuitions.

This kind of concern opens the door to future research on Moral Reasoning examining whether yesterday's post-hoc rationalizations could be the basis for today's moral reasoning (Greene, 2007). For instance, if, after being given the opportunity to explicitly formulate the action principle for themselves, subjects were given a new set of dilemmas involving actions and omissions, would they apply the action principle even more consistently? If so, this would constitute strong evidence that subjects are capable of engaging in principled deontological Reasoning. Likewise, experiments aimed at disrupting the Reasoning process (e.g., using cognitive load or transcranial magnetic stimulation (TMS)) would provide further evidence for this interpretation.

6. Rejecting unprincipled intuitions

Cushman et al. (2006), in addition to examining the action principle, examined the ''contact principle,'' according to which using physical contact to cause harm to a victim (e.g., by pushing) is less morally acceptable than causing harm to a victim without using physical contact (e.g., by dropping the victim through a switch-operated trap door). Subjects' judgments were often consistent with this principle, and, when they were, subjects cited the contact principle 60% of the time. (Note that more recent evidence indicates that these effects are due to the presence of ''personal force'' rather than physical contact per se; Greene, Cushman et al. 2009.) Moreover, 20% of the subjects not only cited the contact principle, but, upon citing it as the basis for their judgments, went on to reject it as a legitimate moral principle, often stating that the mere presence of physical contact did not seem like a good reason to make a moral distinction between the contact and no-contact dilemmas. In other words, subjects revised their judgments when they became conscious of the fact that their initial judgments were *inconsistent* with their beliefs about what kinds of things ought to make a moral difference.

People's willingness to abandon the contact principle is an example of a more general phenomenon in which judgments about pairs of items change depending on whether the items are presented separately or jointly (Bazerman, Loewenstein, & White, 1992). Such effects are typically observed when joint presentation highlights features of the items that are likely to influence judgment, but that strike people, upon reflection, as irrelevant (or weakly relevant) to the judgments being made. For example, when two dictionaries are evaluated separately, subjects are likely to ignore the number of entries in each dictionary, but attend to the fact that one has a slightly torn cover. But when the dictionaries are evaluated jointly, the number of words becomes more important and the fact that one dictionary has a torn cover becomes less important. Joint evaluation makes people think about what they're thinking about and adjust their judgments accordingly.

A recent study (Paharia, Kassam, Greene, & Bazerman, 2009) demonstrates this phenomenon of ''joint/separate reversal'' in a moral context. Subjects responded to scenarios in which a major pharmaceutical company increased its profits by dramatically increasing the price of a slow-selling, but desperately needed, cancer drug. In one scenario, the firm *directly* increased the price of the drug. In the other, the firm increased the price of the drug indirectly by selling the rights to market the drug to a smaller company, knowing that that the other company would increase the price. When the two cases were presented separately (between-subjects), the action in the indirect selling case was judged to be more morally acceptable than the action in the direct selling case. But when the two cases were presented jointly (within-subjects), this effect went away. This is presumably because subjects rejected their intuitive tendency to see direct harm as worse than indirect harm, based on a principled conception of what does and does not matter morally when it comes to evaluating harmful actions.

A study examining the role of disgust in moral judgment (Wheatley & Haidt, 2005) makes a similar point. Highly hypnotizable subjects were given a hypnotic suggestion to feel a ''flash of disgust'' upon hearing affectively neutral words that were embedded within vignettes describing moral violations (e.g., shoplifting, bribery, library theft, etc.). Subjects were asked for judgments concerning how disgusting and how morally acceptable/unacceptable they found each action. When the scenario descriptions included the hypnotic disgust word, subjects judged the actions to be both more disgusting and less morally acceptable, as compared to when the scenario descriptions did not include the hypnotic disgust word. However, the magnitude of the effect was larger for the disgust judgments than for the moral judgments, suggesting that the disgust responses were often countered by some kind of competing response. Wheatley and Haidt also presented subjects with a scenario about a student council representative who finds interesting topics for discussion, doing nothing remotely immoral in the process. Amazingly, some subjects condemned this innocent student when they were hypnotically induced to feel a flash of disgust. Less amazingly, but equally important for present purposes, is the fact that most subjects who were hypnotically induced to experience disgust did not condemn the student council representative. This implies that they overrode their disgust responses, presumably by thinking to themselves something like this: ''I get an icky feeling from this person, but I cannot see how he is doing anything wrong, so I guess his behavior is fine.'' Alternatively, one might propose that this sober response is just as intuitive as the icky feeling. However, even if that is true, this does not explain why subjects so reliably sided with the sober intuition over the icky feeling. This asymmetry in judgment suggests a conscious, principled resolution of the conflict (''But that just makes no sense''). Nevertheless, we cannot rule out the possibility that the sober intuition simply overpowered the icky feeling, without the help of a conscious, principled decision process. As above, studies using fMRI, TMS, or cognitive load could provide support for the hypothesis that such sober judgments in the face of irrational disgust responses are genuinely counter-intuitive and not just differently intuitive.

In short, these studies indicate that people spontaneously reject certain intuitive judgments as unprincipled, so long as they are in a position to appreciate the nature of their intuitions. Moreover, these judgments, at least in some cases, do not appear to be based on countervailing moral intuitions, but rather on more abstract, principled conceptions of what factors ought or ought not carry moral weight.

7. Application of utilitarian moral principles

Consider the following scenario, known as the *crying baby* dilemma (Greene et al., 2001, 2004):

> It's wartime. You and your fellow villagers are hiding from nearby enemy soldiers in a basement. Your baby starts to cry, and you cover your baby's mouth to block the sound. If you remove your hand, your baby will cry loudly, and the soldiers will hear. They will find you, your baby, and the others, and they will kill all of you. If you do not remove your hand, your baby will smother to death. Is it morally acceptable to smother your baby to death in order to save yourself and the other villagers?

Most people find dilemmas such as this difficult, as indicated by relatively long reaction times (RTs) and divergent judgments between subjects. This difficulty appears to be due to a conflict between two competing responses, an automatic emotional response that opposes ''personally'' harmful actions (Greene, Cushman et al. 2009; Greene et al., 2001) and a more controlled cognitive response that, in utilitarian fashion, favors minimizing harm. Evidence for the presence of competing responses comes from an fMRI experiment (Greene et al., 2004) demonstrating that dilemmas such as this one preferentially engage the anterior cingulate cortex (ACC), a brain region known for its role in the detection of response conflict (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Likewise, these difficult dilemmas preferentially engage regions within the dorsolateral prefrontal cortex (DLPFC), which is, once again, known for its role in cognitive control processes that resolve response conflicts (MacDonald et al., 2000; Miller & Cohen, 2001). Critically, these regions of DLPFC also exhibit increased activity when people make utilitarian judgments in response to such dilemmas (as compared to RT-matched nonutilitarian judgments), thus supporting a link between utilitarian judgment and controlled cognitive processing. Likewise, a more recent fMRI study demonstrates that utilitarian judgments approving of breaking a promise in order to save additional lives are also associated with increased DLPFC activity (Greene, Lowenberg, et al. unpublished data).

These results suggest that people sometimes Reason their way to moral judgments by applying utilitarian moral principles. To provide further support for this hypothesis, Greene, Morelli, Lowenberg, Nystrom, and Cohen conducted a study in which subjects responded to dilemmas like the one above while simultaneously engaged in a cognitive load task designed to interfere with controlled cognitive processing. Putting subjects under cognitive load was found to have a selective effect on RT, slowing down utilitarian moral judgments while having no effect on nonutilitarian judgments. In other words, a disruption in one's executive processing makes it harder to render utilitarian judgments, but does not interfere at all with one's ability to render nonutilitarian judgments, a result that holds both for subjects with generally utilitarian inclinations and for subjects with generally nonutilitarian inclinations. This result is consistent with the hypothesis that utilitarian judgments are preferentially supported by ''top-down'' Moral Reasoning processes, and not simply by competing moral intuitions.

However, it should be noted that the load manipulation affected subjects' reaction times, but not their judgments. The reason for this lack of a judgment effect could be traced to a general awareness on the part of the subjects that the load manipulation was causing interference, leading subjects to increase their efforts to overcome the interference. Such a process would be cognitively analogous to the resolution of a speed-accuracy trade-off in favor of accuracy. That is, subjects may avoid making fewer utilitarian judgments by taking longer to make the utilitarian judgments that they make. Like drivers presented with obstacles on the roadway, these subjects may be delayed in, but not prevented from, reaching their destinations.

Three recent behavioral studies provide further evidence for a link between utilitarian judgment and controlled cognition. Hardman (unpublished data) used the Cognitive Reflection Test (CRT), developed by Frederick (2005), to study the relationship between the tendency to override intuitive responses and the tendency to make utilitarian judgments. The CRT asks questions such as the following: ''A bat and a ball cost $1.10. The bat costs one dollar more than the ball. How much does the ball cost?'' Intuitively, the answer seems to most people to be $0.10. However, a bit of reflection reveals that the correct answer is actually $0.05. Subjects who correctly answered questions like this one were twice as likely to give utilitarian responses to the crying baby dilemma and the footbridge dilemma. Likewise, Bartels (2008) found that individuals with more ''rational'' intellectual styles tended to make more utilitarian judgments, while individuals with more ''intuitive'' intellectual styles tended to make fewer utilitarian judgments. Finally, Moore, Clark, and Kane (2008) found that, for a restricted class of dilemmas, individuals with greater working memory capacity tended to make more utilitarian judgments.

Taken together, these studies suggest that utilitarian judgments are preferentially supported by Moral Reasoning. This is because cognitive control generally requires the ''top down'' application of a guiding rule or principle. As noted above, cognitive control mechanisms play a critical role in conforming behavior to the explicit task demands, particularly when the demands of the task are at odds with a prepotent response, as in the Stroop (1935) task (Botvinick et al., 2001; MacDonald et al., 2000). To apply a rule that overrides an intuitive response, one must first determine that the intuitive response is incompatible with the rule, that is, engage in Moral Reasoning. Moreover, it appears that such determinations are conscious. In our research experience, subjects who make utilitarian judgments in response to moral dilemmas invariably justify their answers by appeal to utilitarian principles. These contrasts starkly with the trouble subjects often have in articulating deontological principles to which their judgments conform (Cushman et al., 2006).

8. Overriding implicit negative attitudes

It is well-known that people's implicit attitudes may differ from the explicit attitudes they profess to hold (Greenwald & Banaji, 1995). For example, most White people profess to have either neutral or positive attitudes toward Blacks, and yet a majority of Whites exhibit an implicit anti-Black/pro-White bias (Nosek et al., 2007). This dark cloud, however, may have a silver lining. Since Martin Luther King Jr. uttered the words ''I have a dream…'' in 1963, we have made great strides toward racial equality—not great enough, but very great. Notably, this has occurred *without* eliminating people's implicit racial biases, or even coming close to doing so. A similar phenomenon exists with respect to attitudes toward gays. For example, Inbar and colleagues (Inbar, Pizarro, Knobe, & Bloom, 2009) have shown that college students at a relatively liberal university, whose explicit attitudes toward gays are overwhelmingly nonnegative, exhibit negative attitudes toward gays on implicit measures. Here, too, the glass may be seen as half full: College campuses are tremendously more gay-friendly than they were a generation ago, despite the continued widespread prevalence of negative implicit attitudes toward gays.

How did this happen? One possibility is that, to the extent that we have made social progress within these domains, implicit negative attitudes toward Blacks and other minorities have been outcompeted by other implicit attitudes, such as positive attitudes toward Blacks and gays, positive attitudes toward equality more generally, or negative attitudes toward discrimination. We have no doubt that this is part of the story. But is it the whole story? Could it be that the voices of the civil rights movement, in addition to reshaping people's moral intuitions, succeeded in causing people to *transcend* their intuitions?

Some evidence suggests that social progress is not simply a matter of replacing one dominant intuition with another. For example, Cunningham and colleagues (Cunningham et al., 2004) have shown that when White people view Black faces they exhibit a strong amygdala response when the faces are presented subliminally, but a weaker amygdala response when the faces are presented superliminally. What's more, the superliminal presentations are associated with increased activity in brain regions associated with cognitive control, including the DLPFC. Likewise, interracial interaction is cognitively depleting (as indicated by reduced Stroop (1935) task performance) for Whites who exhibit strong implicit negative associations with Blacks (Richeson & Shelton, 2003). In other words, people do cognitive work to overcome their biases. If this willingness to actively override bias is a product of social influence—and it is hard to imagine that it is not—this suggests that social influence on moral judgment is not simply a matter of changing intuitions.

Suppose, contrary to this interpretation, that the effect of social influence has been nothing more than to implant in people more congenial implicit attitudes, which compete with the nasty old ones. Under this supposition, one might characterize the cognitive control activity observed in these studies as the engagement of a mechanism that mediates between these competing automatic processes. While it is undoubtedly true that these mechanisms play such a mediating role, both here and elsewhere, the critical question is whether these mechanisms serve as a *neutral party* in such mediations. The supposition that these control mechanisms are neutral is hard to reconcile with the fact that explicit measures show no bias: If biased automatic processes are sufficiently powerful to dominate on implicit measures, then why do they have such little effect on explicit measures? The answer, we propose, is that the control mechanisms are not neutral. Rather, their operations reflects a conscious commitment to *principle*, choosing one of two competing automatic responses based on its conformance with that principle, where the favored response is typically the *weaker* of the two automatic responses (as indexed by implicit measures). As noted above, this is exactly parallel to what happens in the color-naming Stroop task: The automatic tendency to read the word (e.g., ''red'' written in blue) is stronger than the competing tendency to name the color of the word, but cognitive control mechanisms nevertheless manage to produce the color-naming behavior most of the time. And this is not because some prior process has made color-naming tendency more potent, but because the cognitive control system is actively conforming the subject's behavior to a ''principle,'' that is, the task instructions to follow the color-naming rule.

Thus, it appears that, when it comes to regulating morally unacceptable bias, cognitive control mechanisms are not merely mediating among competing automatic processes. Rather, it appears that they actively select responses that conform to a rule, a principle such as ''don't discriminate based on race,'' such that the response selected may be less intuitive, less automatic, than its competitors. Moreover, we emphasize that the regulation of bias is not merely a case of private ''reasoned judgment,'' although it may be that as well. This is because the tendency to control one's biased intuitions is (we presume) a tendency that has become widespread due to social influence. Our claim then, is that, in addition to shaping one another's moral intuitions through our words and deeds, we transmit to one another moral principles that allow us to transcend our dominant automatic responses, and thus effect important social change.

9. Conclusion

There are good reasons to believe that people engage in Moral Reasoning and that moralists can influence each other, not simply by modifying each others' intuitions, but by transmitting moral principles that may be used to override moral intuitions, including intuitions that would otherwise dominate behavior. People reject judgments based on their own intuitions when those judgments appear to be unprincipled, particularly if they are given a little prodding (''be rational'') or put in a situation in which they can identify the factors to which their ''unprincipled'' intuitions are sensitive. People appear to spontaneously apply utilitarian moral principles, and perhaps deontological moral principles as well. While much of the evidence for Moral Reasoning comes from laboratory experiments with limited ecological validity, there is some evidence to suggest that Moral Reasoning is a potent social force. When it comes to making moral progress, the ''head'' may be no less indispensible than the ''heart.''[2]

Notes

1. This definition is inspired in part by John Rawls (1971) notion of ''wide-scope reflective equilibrium.'' For a similar account of the psychology of moral reasoning, see Harman, Mason, and Sinnott-Armstrong (in press).
2. For their very helpful comments and insightful suggestions on earlier drafts of this paper, we thank Jonathan Haidt, Bryce Huebner, Joshua Knobe, Molly Pinter, and Wendell Wallach.

References

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

Bazerman, M. H., Loewenstein, G. F., & White, S. B. (1992). Reversals of preference in allocation decisions: Judging an alternative versus choosing among alternatives. *Administrative Science Quarterly*, *37*(2), 220–240.

Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817.

Botvinick, M., Braver, T., Barch, D., Carter, C., & Cohen, J. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652.

Bucciarelli, M., Khemlani, S., & Johnson-Laird, P. N. (2008). The psychology of moral reasoning. *Judgment and Decision Making*, *3*(2), 121–139.

Ciaramelli, E., Muccioli, M., Ladavas, E., & di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Social Cognitive and Affective Neuroscience*, *2*(2), 84.

Cunningham, W. A., Johnson, M. K., Raye, C. L., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2004). Separable neural components in the processing of black and white faces. *Psychological Science*, *15*(12), 806–813.

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.

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10*(5), 204–211.

Foot, P. (1978). *The problem of abortion and the doctrine of double effect*. Oxford, England: Blackwell.

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

Greene, J. D. (2007). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology*, Vol. 3 (pp. 35–79). Cambridge, MA: MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, *44*(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.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27.

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., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628.

Harman, G., Mason, K., & Sinnott-Armstrong, W. (in press). Moral reasoning. In J. M. Doris & the Moral Psychology Research Group (Eds.), *The handbook of moral psychology*. Oxford, England: Oxford University Press.

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. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151–235). New York: Academic Press.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of dorsolateral prefrontal cortex and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1837.

Mendez, M., Anderson, E., & Shapira, J. (2005). An investigation of moral judgement in frontotemporal dementia. *Cognitive and Behavioral Neurology*, *18*(4), 193–197.

Mikhail, J. (2010). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202.

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

Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., Smith, C. T., Olson, K. R., Chugh, D., Greenwald, A. G., & Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology*, *18*(1), 36–88.

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.

Petrinovich, L., O'Neill, P., & Jorgensen, M. (1993). An empirical study of moral intuitions: Toward and evolutionary ethics. *Journal of Personality and Social Psychology*, *64*(3), 467–478.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of moral intuitions: Comment on Haidt (2001). *Psychological Review*, *110*(1), 197–198.

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.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Belknap Press.

Richeson, J. A., & Shelton, J. N. (2003). When prejudice does not pay: Effects of interracial contact on executive function. *Psychological Science*, *14*(3), 287–290.

Smetana, J., & Killen, M. (2006). *Handbook of moral development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Thomson, J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.

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

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