2 The Secret Joke of Kant's Soul

Joshua D. Greene

Two things fill the mind with ever new and increasing wonder and awe, the oftener and more steadily we reflect on them: the starry heavens above me and the moral law within me.

—Immanuel Kant

That such an unnatural use (and so misuse) of one's sexual attributes is a violation of one's duty to himself and is certainly in the highest degree opposed to morality strikes everyone upon his thinking of it . . . However, it is not so easy to produce a rational demonstration of the inadmissibility of that unnatural use, and even the mere unpurposive use, of one's sexual attributes as being a violation of one's duty to himself (and indeed in the highest degree where the unnatural use is concerned). The ground of proof surely lies in the fact that a man gives up his personality (throws it away) when he uses himself merely as a means for the gratification of an animal drive.

—Immanuel Kant, "Concerning Wanton Self-Abuse"

Kant's Joke—Kant wanted to prove, in a way that would dumbfound the common man, that the common man was right: that was the secret joke of this soul. He wrote against the scholars in support of popular prejudice, but for scholars and not for the people.

—Friedrich Nietzsche

There is a substantial and growing body of evidence suggesting that much of what we do, we do unconsciously, and for reasons that are inaccessible to us (Wilson, 2002). In one experiment, for example, people were asked to choose one of several pairs of pantyhose displayed in a row. When asked to explain their preferences, people gave sensible enough answers, referring to the relevant features of the items chosen—superior knit, sheerness, elasticity, etc. However, their choices had nothing to do with such features because the items on display were in fact identical. People simply had a preference for items on the right-hand side of the display (Nisbett

& Wilson, 1977). What this experiment illustrates—and there are many, many such illustrations—is that people make choices for reasons unknown to them and they make up reasonable-sounding justifications for their choices, all the while remaining unaware of their actual motives and subsequent rationalizations.

Jonathan Haidt applies these psychological lessons to the study of moral judgment in his influential paper, "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment" (Haidt, 2001). He argues that for the most part moral reasoning is a post hoc affair: We decide what's right or wrong on the basis of emotionally driven intuitions, and then, if necessary, we make up reasons to explain and justify our judgments. Haidt concedes that some people, some of the time, may actually reason their way to moral conclusions, but he insists that this is not the norm. More important for the purposes of this essay, Haidt does not distinguish among the various approaches to ethics familiar to moral philosophers: consequentialism, deontology, virtue ethics, etc. Rather, his radical thesis is intended, if only implicitly, to apply equally to the adherents of all moral philosophies, though not necessarily well to moral philosophers as a group (Kuhn, 1991).

Jonathan Baron (Baron, 1994), in contrast, draws a psychological distinction between consequentialist and nonconsequentialist judgments, arguing that the latter are especially likely to be made on the basis of heuristics, simple rules of thumb for decision making. Baron, however, does not regard emotion as essential to these heuristic judgments.

In this chapter, I draw on Haidt's and Baron's respective insights in the service of a bit of philosophical psychoanalysis. I will argue that deontological judgments tend to be driven by emotional responses, and that deontological philosophy, rather than being grounded in moral *reasoning*, is to a large extent an exercise in moral *rationalization*. This is in contrast to consequentialism, which, I will argue, arises from rather different psychological processes, ones that are more "cognitive," and more likely to involve genuine moral reasoning. These claims are strictly empirical, and I will defend them on the basis of the available evidence. Needless to say, my argument will be speculative and will not be conclusive. Beyond this, I will argue that if these empirical claims are true, they may have normative implications, casting doubt on deontology as a school of normative moral thought.

Preliminaries

Defining Deontology and Consequentialism

Deontology is defined by its emphasis on moral rules, most often articulated in terms of *rights* and *duties*. Consequentialism, in contrast, is the view that the moral value of an action is in one way or another a function of its consequences alone. Consequentialists maintain that moral decision makers should always aim to produce the best overall consequences for all concerned, if not directly then indirectly. Both consequentialists and deontologists think that consequences are important, but consequentialists believe that consequences are the *only* things that ultimately matter, while deontologists believe that morality both requires and allows us to do things that do not produce the best possible consequences. For example, a deontologist might say that killing one person in order to save several others is wrong, even if doing so would maximize good consequences (S. Kagan, 1997).

This is a standard explanation of what deontology and consequentialism are and how they differ. In light of this explanation, it might seem that my thesis is false *by definition*. Deontology is rule-based morality, usually focused on rights and duties. A deontological judgment, then, is a judgment made out of respect for certain types of moral rules. From this it follows that a moral judgment that is made on the basis of an emotional response simply cannot be a deontological judgment, although it may appear to be one from the outside. Kant himself was adamant about this, at least with respect to his own brand of deontology. He notoriously claimed that an action performed merely out of sympathy and not out of an appreciation of one's duty lacks moral worth (Kant, 1785/1959, chap. 1; Korsgaard, 1996a, chap. 2).

The assumption behind this objection—and as far as I know it has never been questioned previously—is that consequentialism and deontology are, first and foremost, moral philosophies. It is assumed that philosophers know exactly what deontology and consequentialism are because these terms and concepts were defined by philosophers. Despite this, I believe it is possible that philosophers do not necessarily know what consequentialism and deontology really are.

How could this be? The answer, I propose, is that the terms "deontology" and "consequentialism" refer to *psychological natural kinds*. I believe that consequentialist and deontological views of philosophy are not so much philosophical inventions as they are philosophical manifestations of two dissociable psychological patterns, two different ways of moral thinking,

that have been part of the human repertoire for thousands of years. According to this view, the moral philosophies of Kant, Mill, and others are just the explicit tips of large, mostly implicit, psychological icebergs. If that is correct, then philosophers may not really know what they're dealing with when they trade in consequentialist and deontological moral theories, and we may have to do some science to find out.

An analogy, drawing on a familiar philosophical theme: Suppose that in a certain tropical land the inhabitants refer to water by this symbol: ♦. And in their *Holy Dictionary* it clearly states that ♦ is a clear and drinkable liquid. (That is, the dictionary defines ♦ in terms of its "primary intension" (Chalmers, 1996).) One day an enterprising youngster journeys to the top of a nearby mountain and is the first of her people to encounter ice. Through a bit of experimentation, she discovers that ice is a form of water and excitedly tells the tribal Elders of her discovery. The next day she drags one of the Elders to the mountaintop, hands him some ice, and says, "Behold! ♦!" At which point the exasperated Elder explains that ♦ is a liquid, that the hard stuff in his hand is clearly not a liquid, and that he doesn't appreciate having his time wasted.

In a narrow sense the Elder is correct. The *Holy Dictionary* is the authority on what the local symbols mean, and it states clearly that ♦ refers to a clear, drinkable, liquid. But the Elder is missing the big picture. What he is forgetting, or perhaps never understood, is that many things in the world have underlying structures—"essences," if you prefer—that are responsible for making things appear and behave as they do, for giving them their functional properties. And because things have underlying structures, it is possible to refer to something, even make up a definition for it, without really understanding what it is (Kripke, 1980; Putnam, 1975). Of course, a linguistic community can insist that their definition is correct. No one's to stop them from using their symbols as they please. However, in doing this, they run the risk of missing the big picture, of denying themselves a deeper understanding of what's going on around them, or even within them.

Because I am interested in exploring the possibility that deontology and consequentialism are psychological natural kinds, I will put aside their conventional philosophical definitions and focus instead on their relevant functional roles. As noted earlier, consequentialists and deontologists have some characteristic practical disagreements. For example, consequentialists typically say that killing one person in order to save several others may be the right thing to do, depending on the situation. Deontologists, in contrast, typically say that it's wrong to kill one person for the benefit of

others, that the "ends don't justify the means." Because consequentialists and deontologists have these sorts of practical disagreements, we can use these disagreements to define consequentialist and deontological judgments functionally. For the purposes of this discussion, we'll say that consequentialist judgments are judgments in favor of characteristically consequentialist conclusions (e.g., "Better to save more lives") and that deontological judgments are judgments in favor of characteristically deontological conclusions (e.g., "It's wrong despite the benefits"). My use of "characteristically" is obviously loose here, but I trust that those familiar with contemporary ethical debates will know what I mean. Note that the kind of judgment made is largely independent of who is making it. A card-carrying deontologist can make a "characteristically consequentialist" judgment, as when Judith Jarvis Thomson says that it's okay to turn a runaway trolley that threatens to kill five people onto a side track so that it will kill only one person instead (Thomson, 1986). This is a "characteristically consequentialist" judgment because it is easily justified in terms of the most basic consequentialist principles, while deontologists need to do a lot of fancy philosophizing in order to defend this position. Likewise, consider the judgment that it's wrong to save five people who need organ transplants by removing the organs from an unwilling donor (Thomson, 1986). This judgment is "characteristically deontological," not because many card-carrying consequentialists don't agree, but because they have to do a lot of extra explaining to justify their agreement.

By defining "consequentialism" and "deontology" in terms of their characteristic judgments, we give our empirical hypothesis a chance. If it turns out that characteristically deontological judgments are driven by emotion (an empirical possibility), then that raises the possibility that deontological *philosophy* is also driven by emotion (a further empirical possibility). In other words, what we find when we explore the psychological causes of characteristically deontological judgments might suggest that what deontological moral philosophy really is, what it is *essentially*, is an attempt to produce rational justifications for emotionally driven moral judgments, and not an attempt to reach moral conclusions on the basis of moral reasoning.

The point for now, however, is simply to flag the terminological issue. When I refer to something as a "deontological judgment" I am saying that it is a characteristically deontological judgment and am not insisting that the judgment in question necessarily meets the criteria that philosophers would impose for counting that judgment as deontological. In the end, however, I will argue that such judgments are best understood as genuinely

deontological because they are produced by an underlying psychology that is the hidden essence of deontological philosophy.

Defining "Cognition" and "Emotion"

In what follows I will argue that deontological judgment tends to be driven by emotion, while consequentialist judgment tends to be driven by "cognitive" processes. What do we mean by "emotion" and "cognition," and how do these things differ?

Sometimes "cognition" refers to information processing in general, as in "cognitive science," but often "cognition" is used in a narrower sense that contrasts with "emotion," despite the fact that emotions involve information processing. I know of no good off-the-shelf definition of "cognition" in this more restrictive sense, despite its widespread use. Elsewhere, my collaborators and I offered a tentative definition of our own (Greene, Nystrom, Engell, Darley, & Cohen, 2004), one that is based on the differences between the information-processing requirements of stereotyped versus flexible behavior.

The rough idea is that "cognitive" representations are inherently neutral representations, ones that do not automatically trigger particular behavioral responses or dispositions, while "emotional" representations do have such automatic effects, and are therefore behaviorally valenced. (To make things clear, I will use quotation marks to indicate the more restrictive sense of "cognitive" defined here, and I will drop the quotation marks when using this term to refer to information processing in general.) Highly flexible behavior requires "cognitive" representations that can be easily mixed around and recombined as situational demands vary, and without pulling the agent in sixteen different behavioral directions at once. For example, sometimes you need to avoid cars, and other times you need to approach them. It is useful, then, if you can represent CAR in a behaviorally neutral or "cognitive" way, one that doesn't automatically presuppose a particular behavioral response. Stereotyped behavior, in contrast, doesn't require this sort of flexibility and therefore doesn't require "cognitive" representations, at least not to the same extent.

While the whole brain is devoted to cognition, "cognitive" processes are especially important for reasoning, planning, manipulating information in working memory, controlling impulses, and "higher executive functions" more generally. Moreover, these functions tend to be associated with certain parts of the brain, primarily the dorsolateral surfaces of the prefrontal cortex and parietal lobes (Koechlin, Ody, & Kouneiher, 2003; Miller & Cohen, 2001; Ramnani & Owen, 2004). Emotion, in contrast, tends to

be associated with other parts of the brain, such as the amygdala and the medial surfaces of the frontal and parietal lobes (Adolphs, 2002; Maddock, 1999; Phan, Wager, Taylor, & Liberzon, 2002). And while the term "emotion" can refer to stable states such as moods, here we will primarily be concerned with emotions subserved by processes that in addition to being valenced, are quick and automatic, though not necessarily conscious.

Here we are concerned with two different kinds of moral judgment (deontological and consequentialist) and two different kinds of psychological process ("cognitive" and emotional). Crossing these, we get four basic empirical possibilities. First, it could be that both kinds of moral judgment are generally "cognitive," as Kohlberg's theories suggest (Kohlberg, 1971).[1] At the other extreme, it could be that both kinds of moral judgment are primarily emotional, as Haidt's view suggests (Haidt, 2001). Then there is the historical stereotype, according to which consequentialism is more emotional (emerging from the "sentimentalist" tradition of David Hume (1740/1978) and Adam Smith (1759/1976)) while deontology is more "cognitive" [encompassing the Kantian "rationalist" tradition (Kant, 1959)]. Finally, there is the view for which I will argue, that deontology is more emotionally driven while consequentialism is more "cognitive." I hasten to add, however, that I don't believe that either approach is strictly emotional or "cognitive" (or even that there is a sharp distinction between "cognition" and emotion). More specifically, I am sympathetic to Hume's claim that all moral judgment (including consequentialist judgment) must have some emotional component (Hume, 1978). But I suspect that the kind of emotion that is essential to consequentialism is fundamentally different from the kind that is essential to deontology, the former functioning more like a currency and the latter functioning more like an alarm. We will return to this issue later.

Scientific Evidence

Evidence from Neuroimaging

In recent decades, philosophers have devised a range of hypothetical moral dilemmas that capture the tension between the consequentialist and deontological viewpoints. A well-known handful of these dilemmas gives rise to what is known as the "trolley problem" (Foot, 1978; Thomson, 1986), which begins with the *trolley* dilemma.

A runaway trolley is headed for five people who will be killed if it proceeds on its present course. The only way to save these people is to hit a

switch that will turn the trolley onto a side track, where it will run over and kill one person instead of five. Is it okay to turn the trolley in order to save five people at the expense of one? The consensus among philosophers (Fischer & Ravizza, 1992), as well as people who have been tested experimentally (Petrinovich & O'Neill, 1996; Petrinovich, O'Neill, & Jorgensen, 1993), is that it is morally acceptable to save five lives at the expense of one in this case.

Next consider the *footbridge* dilemma (Thomson, 1986): As before, a runaway trolley threatens to kill five people, but this time you are standing next to a large stranger on a footbridge spanning the tracks, in between the oncoming trolley and the five people. The only way to save the five people is to push this stranger off the bridge and onto the tracks below. He will die as a result, but his body will stop the trolley from reaching the others. Is it okay to save the five people by pushing this stranger to his death? Here the consensus is that it is not okay to save five lives at the expense of one (Fischer & Ravizza, 1992; Greene et al., 2004; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Petrinovich & O'Neill, 1996; Petrinovich et al., 1993).

People exhibit a characteristically consequentialist response to the *trolley* case and a characteristically deontological response to the *footbridge* case. Why? Philosophers have generally offered a variety of *normative* explanations. That is, they have assumed that our responses to these cases are correct, or at least reasonable, and have sought principles that *justify* treating these two cases differently (Fischer & Ravizza, 1992). For example, one might suppose, following Kant (1785/1959) and Aquinas (1265–1272/1988), that it is wrong to harm someone as a means to helping someone else. In the *footbridge* case the proposed action involves literally using the person on the footbridge as a trolley stopper, whereas in the *trolley* case the victim is to be harmed merely as a side effect. (Were the single person on the alternative track to magically disappear, we would be very pleased.) In response to this proposal, Thomson devised the *loop* case (Thomson, 1986). Here the situation is similar to that of the *trolley* dilemma, but this time the single person is on a piece of track that branches off of the main track and then rejoins it at a point before the five people. In this case, if the person were not on the side track, the trolley would return to the main track and run over the five people. The consensus here is that it is morally acceptable to turn the trolley in this case, despite the fact that here, as in the *footbridge* case, a person will be used as a means.

There have been many such normative attempts to solve the trolley problem, but none of them has been terribly successful (Fischer & Ravizza,

1992). My collaborators and I have proposed a partial and purely descriptive solution to this problem and have collected some scientific evidence in favor of it. We hypothesized that the thought of pushing someone to his death in an "up close and personal" manner (as in the *footbridge* dilemma) is more emotionally salient than the thought of bringing about similar consequences in a more impersonal way (e.g., by hitting a switch, as in the *trolley* dilemma). We proposed that this difference in emotional response explains why people respond so differently to these two cases. That is, people tend toward consequentialism in the case in which the emotional response is low and tend toward deontology in the case in which the emotional response is high.

The rationale for distinguishing between *personal* and *impersonal* forms of harm is largely evolutionary. "Up close and personal" violence has been around for a very long time, reaching far back into our primate lineage (Wrangham & Peterson, 1996). Given that personal violence is evolutionarily ancient, predating our recently evolved human capacities for complex abstract reasoning, it should come as no surprise if we have innate responses to personal violence that are powerful but rather primitive. That is, we might expect humans to have negative emotional responses to certain basic forms of interpersonal violence, where these responses evolved as a means of regulating the behavior of creatures who are capable of intentionally harming one another, but whose survival depends on cooperation and individual restraint (Sober & Wilson, 1998; Trivers, 1971). In contrast, when a harm is *impersonal*, it should fail to trigger this alarmlike emotional response, allowing people to respond in a more "cognitive" way, perhaps employing a cost-benefit analysis. As Josef Stalin once said, "A single death is a tragedy; a million deaths is a statistic." His remarks suggest that when harmful actions are sufficiently impersonal, they fail to push our emotional buttons, despite their seriousness, and as a result we think about them in a more detached, actuarial fashion.

This hypothesis makes some strong predictions regarding what we should see going on in people's brains while they are responding to dilemmas involving personal versus impersonal harm (henceforth called "personal" and "impersonal" moral dilemmas). The contemplation of personal moral dilemmas like the *footbridge* case should produce increased neural activity in brain regions associated with emotional response and social cognition, while the contemplation of impersonal moral dilemmas like the *trolley* case should produce relatively greater activity in brain regions associated with "higher cognition."[2] This is exactly what was observed (Greene et al., 2004; Greene et al., 2001). Contemplation of personal moral dilemmas produced

relatively greater activity in three emotion-related areas: the posterior cingulate cortex, the medial prefrontal cortex, and the amygdala. This effect was also observed in the superior temporal sulcus, a region associated with various kinds of social cognition in humans and other primates (Allison, Puce, & McCarthy, 2000; Saxe, Carey, & Kanwisher, 2004a). At the same time, contemplation of impersonal moral dilemmas produced relatively greater neural activity in two classically "cognitive" brain areas, the dorsolateral prefrontal cortex and inferior parietal lobe.

This hypothesis also makes a prediction regarding people's reaction times. According to the view I have sketched, people tend to have emotional responses to personal moral violations, responses that incline them to judge against performing those actions. That means that someone who judges a personal moral violation to be *appropriate* (e.g., someone who says it's okay to push the man off the bridge in the *footbridge* case) will most likely have to override an emotional response in order to do it. This overriding process will take time, and thus we would expect that "yes" answers will take longer than "no" answers in response to personal moral dilemmas like the *footbridge* case. At the same time, we have no reason to predict a difference in reaction time between "yes" and "no" answers in response to impersonal moral dilemmas like the *trolley* case because there is, according to this model, no emotional response (or much less of one) to override in such cases. Here, too, the prediction has held. Trials in which the subject judged in favor of personal moral violations took significantly longer than trials in which the subject judged against them, but there was no comparable reaction time effect observed in response to impersonal moral violations (Greene et al., 2004; Greene et al., 2001).

Further results support this model as well. Next we subdivided the personal moral dilemmas into two categories on the basis of difficulty (i.e., based on reaction time). Consider the following moral dilemma (the *crying baby* dilemma): It is wartime, and you and some of your fellow villagers are hiding from 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, the soldiers will hear, and they will find you and the others and kill everyone they find, including you and your baby. If you do not remove your hand, your baby will smother to death. Is it okay to smother your baby to death in order to save yourself and the other villagers?

This is a very difficult question. Different people give different answers, and nearly everyone takes a relatively long time. This is in contrast to other personal moral dilemmas, such as the *infanticide* dilemma, in which a

teenage girl must decide whether to kill her unwanted newborn. In response to this case, people (at least the ones we tested) quickly and unanimously say that this action is wrong.

What's going on in these two cases? My colleagues and I hypothesized as follows. In both cases there is a prepotent, negative emotional response to the personal violation in question, killing one's own baby. In the *crying baby* case, however, a cost-benefit analysis strongly favors smothering the baby. After all, the baby is going to die no matter what, and so you have nothing to lose (in consequentialist terms) and much to gain by smothering it, awful as it is. In some people the emotional response dominates, and those people say "no." In other people, this "cognitive," cost-benefit analysis wins out, and these people say "yes."

What does this model predict that we will see going on in people's brains when we compare cases like *crying baby* and *infanticide*? First, this model supposes that cases like *crying baby* involve an increased level of "response conflict," that is, conflict between competing representations for behavioral response. Thus, we should expect that difficult moral dilemmas like *crying baby* will produce increased activity in a brain region that is associated with response conflict, the anterior cingulate cortex (Botvinick, Braver, Barch, Carter, & Cohen, 2001). Second, according to our model, the crucial difference between cases like *crying baby* and those like *infanticide* is that the former evoke strong "cognitive" responses that can effectively compete with a prepotent, emotional response. Thus, we should expect to see increased activity in classically "cognitive" brain areas when we compare cases like *crying baby* with cases like *infanticide*, despite the fact that difficult dilemmas like *crying baby* are personal moral dilemmas, which were previously associated with emotional response (Greene et al., 2001).

These two predictions have held (Greene et al., 2004). Comparing high-reaction-time personal moral dilemmas like *crying baby* with low-reaction-time personal moral dilemmas like *infanticide* revealed increased activity in the anterior cingulate cortex (conflict) as well as the anterior dorsolateral prefrontal cortex and the inferior parietal lobes, both classically "cognitive" brain regions.

Cases like *crying baby* are especially interesting because they allow us to directly compare the neural activity associated with characteristically consequentialist and deontological responses. According to our model, when people say "yes" to such cases (the consequentialist answer), it is because the "cognitive" cost-benefit analysis has successfully dominated the prepotent emotional response that drives people to say "no" (the deontological answer). If that is correct, then we should expect to see increased

activity in the previously identified "cognitive" brain regions (the dorsolateral prefrontal cortex and inferior parietal cortex) for the trials in which people say "yes" in response to cases like *crying baby*. This is exactly what we found. In other words, people exhibit more "cognitive" activity when they give the consequentialist answer.[3]

To summarize, people's moral judgments appear to be products of at least two different kinds of psychological processes. First, both brain imaging and reaction-time data suggest that there are prepotent negative emotional responses that drive people to disapprove of the personally harmful actions proposed in cases like the *footbridge* and *crying baby* dilemmas. These responses are characteristic of deontology, but not of consequentialism. Second, further brain imaging results suggest that "cognitive" psychological processes can compete with the aforementioned emotional processes, driving people to approve of personally harmful moral violations, primarily when there is a strong consequentialist rationale for doing so, as in the *crying baby* case. The parts of the brain that exhibit increased activity when people make characteristically consequentialist judgments are those that are most closely associated with higher cognitive functions such as executive control (Koechlin et al., 2003; Miller and Cohen, 2001), complex planning (Koechlin, Basso, Pietrini, Panzer, & Grafman, 1999), deductive and inductive reasoning (Goel & Dolan, 2004), taking the long view in economic decision making (McClure, Laibson, Loewenstein, & Cohen., 2004), and so on. Moreover, these brain regions are among those most dramatically expanded in humans compared with other primates (Allman, Hakeem, & Watson, 2002).

Emotion and the Sense of Moral Obligation

In his classic article, "Famine, Affluence, and Morality," Peter Singer (1972) argues that we in the affluent world have an obligation to do much more than we do to improve the lives of needy people. He argues that if we can prevent something very bad from happening without incurring a comparable moral cost, then we ought to do it. For example, if one notices a small child drowning in a shallow pond, one is morally obliged to wade in and save that child, even if it means muddying one's clothes. As Singer points out, this seemingly innocuous principle has radical implications, implying that all of us who spend money on unnecessary luxuries should give up those luxuries in order to spend the money on saving and/or improving the lives of impoverished peoples. Why, Singer asks, do we have a strict obligation to save a nearby drowning child but no comparable

obligation to save faraway sick and starving children through charitable donations to organizations like Oxfam?

Many normative explanations come to mind, but none is terribly compelling. Are we allowed to ignore the plight of faraway children because they are citizens of foreign nations? If so, then would it be acceptable to let the child drown, provided that the child was encountered while traveling abroad? Or in international waters? And what about the domestic poor? This argument does not relieve us of our obligations to them. Is it because of diffused responsibility—because many are in a position to help a starving child abroad, but only you are in a position to help this hypothetical drowning child? What if there were many people standing around the pond doing nothing? Would that make it okay for you to do nothing as well? Is it because international aid is ultimately ineffective, only serving to enrich corrupt politicians or create more poor people? In that case, our obligation would simply shift to more sophisticated relief efforts incorporating political reform, economic development, family planning education, and so on. Are all relief efforts doomed to ineffectiveness? That is a bold empirical claim that no one can honestly make with great confidence.

Here we find ourselves in a position similar to the one we faced with the trolley problem. We have a strong intuition that two moral dilemmas are importantly different, and yet we have a hard time explaining what that important difference is (S. Kagan, 1989; Unger, 1996). It turns out that the same psychological theory that makes sense of the trolley problem can make sense of Singer's problem. Note that the interaction in the case of the drowning child is "up close and personal," the sort of situation that might have been encountered by our human and primate ancestors. Likewise, note that the donation case does not "up close and personal," and is not the sort of situation that our ancestors could have encountered. At no point were our ancestors able to save the lives of anonymous strangers through modest material sacrifices. In light of this, the psychological theory presented here suggests that we are likely to find the obligation to save the drowning child more pressing simply because that "up close and personal" case pushes our emotional buttons in a way that the more impersonal donation case does not (Greene, 2003). As it happens, these two cases were among those tested in the brain imaging study described earlier, with a variation on the drowning child case included in the *personal* condition and the donation case included in the *impersonal* condition (Greene et al., 2004; Greene et al., 2001).

Few people accept Singer's consequentialist conclusion. Rather, people tend to believe, in a characteristically deontological way, that they are within their moral rights in spending their money on luxuries for themselves, despite the fact that their money could be used to dramatically improve the lives of other people. This is exactly what one would expect if (1) the deontological sense of obligation is driven primarily by emotion, and (2) when it comes to obligations to aid, emotions are only sufficiently engaged when those to whom we might owe something are encountered (or conceived of) in a personal way.

Emotion and the Pull of Identifiable Victims

One aspect of someone's being "up close and personal" is that such a person is always, in some sense, an identifiable, determinate individual and not a mere statistical someone (Greene and Haidt, 2002; Greene et al., 2001). The drowning child, for example, is presented as a particular person, while the children you might help through donations to Oxfam are anonymous and, as far as you know, indeterminate.[4] Many researchers have observed a tendency to respond with greater urgency to identifiable victims, compared with indeterminate, "statistical" victims (Schelling, 1968). This is known as the "identifiable victim effect."

You may recall, for example, the case of Jessica McClure, a.k.a. "Baby Jessica," who in 1987 was trapped in a well in Texas. More than $700,000 was sent to her family to support the rescue effort (Small & Loewenstein, 2003; Variety, 1989). As Small and Loewenstein point out, that amount of money, if it had been spent on preventive healthcare, could have been used to save the lives of many children. This observation raises a normative question that is essentially the same as Singer's. Do we have a greater obligation to help people like Baby Jessica than we do to help large numbers of others who could be saved for less? If all else is equal, a consequentialist would say "no," while, most people apparently would say "yes." Furthermore, most people, if pressed to explain their position, would probably do so in deontological terms. That is, they would probably say that we have a *duty* to aid someone like Baby Jessica, even if doing so involves great effort and expense, while we have no comparable duty to the countless others who might be helped using the same resources.

The same "up close and personal" theory of emotional engagement can explain this pattern of judgment. Others have proposed what amounts to the same hypothesis, and others still have gathered independent evidence to support it. In Thomas Schelling's seminal article on this topic he observes that the death of a particular person invokes "anxiety and sentiment, guilt

and awe, responsibility and religion, [but] . . . most of this awesomeness disappears when we deal with statistical death" (Schelling, 1968; Small & Loewenstein, 2003). Inspired by Schelling's observation, Small and Loewenstein conducted two experiments aimed at testing the hypothesis that "identifiable victims stimulate a more powerful emotional response than do statistical victims."

Their crucial move was to design their experiments in such a way that their results could count against all normative explanations of the identifiable victim effect, i.e., explanations that credit decision makers with normatively respectable reasons for favoring identifiable victims. This is difficult because the process of identifying a victim inevitably provides information about that victim (name, age, gender, appearance, etc.) that could serve as a rational basis for favoring that person. To avoid this, they sought to document a weaker form of the identifiable victim effect, which one might call the "determinate victim effect." They examined people's willingness to benefit determined versus undetermined individuals under conditions in which all meaningful information about the victims was held constant.

Their first experiment worked as follows. Ten laboratory subjects were each given an "endowment" of $10. Some subjects randomly drew cards that said "KEEP" and were allowed to retain their endowments, while other subjects drew cards that said "LOSE" and subsequently had their endowments taken away, thus rendering them "victims." Each of the nonvictim subjects was anonymously paired with one of the victims as a result of drawing that victim's number. The nonvictim subjects were allowed to give a portion of their endowments to their respective victims, and each could choose how much to give. However—the crucial manipulation—some nonvictim subjects drew the victim's number *before* deciding how much to give, while others drew the victim's number *after* deciding, knowing in advance that they would do so later. In other words, some subjects had to answer the question, "How much do I want to give to person #4?" (determined victim), whereas other subjects had to answer the question, "How much do I want to give to the person whose number I will draw?" (undetermined victim). At no point did the nonvictim subjects ever know who would receive their money. The results: The mean donation for the group who gave to determined victims was 60 percent higher than that of the group giving to undetermined victims. The median donation for the determined victim group was more than twice as high.

It is worth emphasizing the absurdity of this pattern of behavior. There is no rational basis for giving more money to "randomly determined

person #4" than to "person #? to be randomly determined," and yet that is what these people did.[5] (Note that the experiment was designed so that none of the participants would ever know who chose what.) Why would people do this? Here, too, the answer implicates emotion. In a follow-up study replicating this effect, the subjects reported on the levels of sympathy and pity they felt for the determined and undetermined victims with whom they were paired. As expected, their reported levels of sympathy and pity tracked their donation levels (Small, personal communication 2/12/05).

One might wonder whether this pattern holds up outside the lab. To find out, Small and Loewenstein conducted a subsequent study in which people could donate money to Habitat for Humanity to provide a home for a needy family, where the family was either determined or to be determined. As predicted, the mean donation was 25 percent higher in the determined family condition, and the median donation in the determined family condition was double that of the undetermined family condition.

And then there is Baby Jessica. We can't say for sure that resources were directed to her instead of to causes that could use the money more effectively because of people's emotional responses (and not because of people's deontological reasoning about rights and duties), but what evidence there is suggests that that is the case. As Stalin might have said, "A determinate individual's death is a tragedy; a million indeterminate deaths is a statistic."

Anger and Deontological Approaches to Punishment

While consequentialists and deontologists agree that punishment of wrongdoing is necessary and important, they disagree sharply over the proper justification for punishment. Consequentialists such as Jeremy Bentham (Bentham, 1789/1982) argue that punishment is justified solely by its future beneficial effects, primarily through deterrence and (in the case of criminal law) the containment of dangerous individuals. While few would deny that the prevention of future harm provides *a* legitimate justification for punishment, many believe that such pragmatic considerations are not the *only* legitimate reasons to punish, or even the main ones. Deontologists such as Kant (1796/2002), for example, argue that the primary justification for punishment is *retribution*, to give wrongdoers what they deserve based on what they have done, regardless of whether such retribution will prevent future wrongdoing.

One might wonder, then, about the psychology of the typical punisher. Do people punish, or endorse punishment, because of its beneficial effects,

or do people punish because they are motivated to give people what they deserve, in proportion to their "internal wickedness," to use Kant's phrase (Carlsmith, Darley, & Robinson, 2002; Kant, 1796–97/2002). Several studies speak to this question, and the results are consistent. People endorse both consequentialist and retributivist justifications for punishment in the abstract, but in practice, or when faced with more concrete hypothetical choices, people's motives appear to be predominantly retributivist. Moreover, these retributivist inclinations appear to be emotionally driven. People punish in proportion to the extent that transgressions make them angry.

First, let us consider whether punitive judgments are predominantly consequentialist or deontological and retributivist.[6] Jonathan Baron and colleagues have conducted a series of experiments demonstrating that people's punitive judgments are, for the most part, retributivist rather than consequentialist. In one study Baron and Ritov (1993) presented people with hypothetical corporate liability cases in which corporations could be required to pay fines. In one set of cases a corporation that manufactures vaccines is being sued because a child died as a result of taking one of its flu vaccines. Subjects were given multiple versions of this case. In one version, it was stipulated that a fine would have a positive deterrent effect. That is, a fine would make the company produce a safer vaccine. In a different version, it was stipulated that a fine would have a "perverse" effect. Instead of causing the firm to make a safer vaccine available, a fine would cause the company to stop making this kind of vaccine altogether, a bad result given that the vaccine in question does more good than harm and that no other firm is capable of making such a vaccine. Subjects indicated whether they thought a punitive fine was appropriate in either of these cases and whether the fine should differ between these two cases. A majority of subjects said that the fine should not differ at all. Baron and Ritov achieved similar results using a complementary manipulation concerning deterrent effects on the decisions of other firms. In a different set of studies Baron and colleagues found a similar indifference to consequentialist factors in response to questions about the management of hazardous waste (Baron, Gowda, & Kunreuther, 1993).

The results of these studies are surprising in light of the fact that many people regard the deterrence of future harmful decisions as a major reason, if not the primary reason, for imposing such fines in the real world. The strength of these results is also worth emphasizing. The finding here is not simply that people's punitive judgments fail to accord with consequentialism, the view that consequences are ultimately the *only*

things that should matter to decision makers. Much more than that, it seems that a majority of people give *no weight whatsoever* to factors that are of clear consequentialist importance, at least in the contexts under consideration.

If people do not punish for consequentialist reasons, what motivates them? In a study by Kahneman and colleagues (Kahneman, Schkade, & Sunstein, 1998), subjects responded to a number of similar hypothetical scenarios (e.g., a case of anemia due to benzene exposure at work). For each scenario subjects rated the extent to which the defendant's action was "outrageous." They also rated the extent to which the defendant in each case should be punished. The correlation between the mean outrage ratings for these scenarios and their mean punishment ratings were nearly perfect, with a Pearson's correlation coefficient (r) of 0.98. (A value of 1 indicates a perfect correlation.) Kahneman and colleagues conclude that the extent to which people desire to see a corporation punished for its behavior is almost entirely a function of the extent to which they are emotionally outraged by that corporation's behavior.

Carlsmith and colleagues (Carlsmith et al., 2002) conducted a similar set of studies aimed explicitly at determining whether people punish for consequentialist or deontological reasons. Here, as earlier, subjects were presented with scenarios involving morally and legally culpable behavior, in this case perpetrated by individuals rather than corporations. As before, subjects were asked to indicate how severe each person's punishment should be, first in abstract terms ("not at all severe" to "extremely severe") and then in more concrete terms ("not guilty"/no punishment to "life sentence"). The experimenters varied the scenarios in ways that warranted different levels of punishment, depending on the rationale for punishment. For example, a consequentialist theory of punishment considers the detection rate associated with a given kind of crime and the publicity associated with a given kind of conviction to be relevant factors in assigning punishments. According to consequentialists, if a crime is difficult to detect, then the punishment for that crime ought to be made more severe in order to counterbalance the temptation created by the low risk of getting caught. Likewise, if a conviction is likely to get a lot of publicity, then a law enforcement system interested in deterrence should take advantage of this circumstance by "making an example" of the convict with a particularly severe punishment, thus getting a maximum of deterrence "bang" for its punishment "buck."

The results were clear. For the experimental group as a whole, there was no significant change in punishment recommendations when the detec-

tion rates and levels of publicity were manipulated. In other words, people were generally indifferent to factors that according to consequentialists should matter, at least to some extent. This is in spite of the fact that Carlsmith et al., as well as others (Weiner, Graham, & Reyna, 1997), found that subjects readily expressed a general kind of support for deterrence-oriented penal systems and corporate policies.

In a follow-up study, subjects were explicitly instructed to adopt a consequentialist approach, with the consequentialist rationale explicitly laid out and with extra manipulation checks included to ensure that the subjects understood the relevant facts. Here, too, the results were striking. Subjects did modify their judgments when they were told to think like consequentialists, but not in a genuinely consequentialist way. Instead of becoming selectively sensitive to the factors that increase the consequentialist benefits of punishment, subjects indiscriminately ratcheted up the level of punishment in all cases, giving perpetrators the punishment that they thought the perpetrators deserved based on their actions, plus a bit more for the sake of deterrence.

What motivated these subjects' punitive judgments? Here, too, an important part of the answer appears to be "outrage." Subjects indicated the extent to which they were "morally outraged" by the offenses in question, and the extent of moral outrage in response to a given offense was a pretty good predictor of the severity of punishment assigned to the perpetrator, although the effect here was weaker than that observed in Kahneman et al.'s study.[7] Moreover, a structural equation model of these data suggests that the factors that had the greatest effect on people's judgments about punishment (severity of the crime, presence of mitigating circumstances) worked their effects through "moral outrage."

You will recall Small and Loewenstein's research on the "identifiable victim effect" discussed in the previous section. More recently they have documented a parallel effect in the domain of punishment. Subjects played an "investment game" in which individuals were given money that they could choose to put into a collective investment pool. The game allows individuals to choose the extent to which they will play cooperatively, benefiting the group at the chooser's expense. After the game, cooperators were given the opportunity to punish selfish players by causing them to lose money, but the punishing cooperators had to pay for the pleasure. As before, the crucial manipulation was between determined and undetermined individuals, in this case the selfish players. Some subjects were asked, "How much would you like to punish uncooperative subject #4?" while others were asked, "How much would you like to punish the

uncooperative subject whose number you will draw?" Consistent with previous results, the average punishment was almost twice as high for the determined group, and once again the subjects' reports of their emotional responses (in this case a composite measure of anger and blame) tracked their behavior (Small & Loewenstein, 2005).

Recent neuroimaging studies also suggest that the desire to punish is emotionally driven. Alan Sanfey, Jim Rilling, and colleagues (Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003) conducted a brain imaging study of the ultimatum game to study the neural bases of people's sense of fairness. The ultimatum game works as follows. There is a sum of money, say $10, and the first player (the proposer) makes a proposal on how to divide it up between her or himself and the other player. The second player, the responder, can either accept the offer, in which case the money is divided as proposed, or reject the offer, in which case no one gets anything. Proposers usually make offers that are fair (i.e., a fifty-fifty split) or close to fair, and responders tend to reject offers that are more than a little unfair. In other words, responders will typically pay for the privilege of punishing unfair proposers, even when the game is played only once. Why do people do this?

The answer, once again, implicates emotion. The experimenters found that unfair offers, compared with fair offers, produced increased activity in the anterior insula, a brain region associated with anger, disgust, and autonomic arousal. Moreover, individuals' average levels of insula activity correlated positively with the percentage of offers they rejected and was weaker for trials in which the subject believed that the unfair offer was made by a computer program rather than a real person. Of course, it is conceivable that people were punishing in an attempt to deter unfair proposers from being unfair to others in the future, but that seems unlikely given the consistent finding that people are insensitive to manipulations that modulate the deterrent effects of punishment. Instead, it seems much more likely that people inflicted punishment for its own sake. And once again, it seems that this retributivist tendency is emotionally driven. A more recent neuroimaging study of punishment in response to violations of trust yields a similar conclusion (de Quervain, Fischbacher, Treyer, Schellhammer, Schnyder, et al., 2004). In this study, the extent of punishment was correlated with the level of activity in the caudate nucleus, a brain region associated with emotion and related more specifically to motivation and reward.

When people are asked in a general and abstract way about why it makes sense to punish, consequentialist arguments are prominent (Carlsmith

et al., 2002; Weiner et al., 1997). However, when people are presented with more concrete cases involving specific individuals carrying out specific offenses, people's judgments are largely, and in many cases completely, insensitive to factors affecting the consequences of punishment. This is so even when the consequentialist rationale for responding to these factors is highlighted and when people are explicitly instructed to think like consequentialists. It seems, then, that consequentialist thinking plays a negligible role in commonsense punitive judgment and that commonsense punitive judgment is almost entirely retributivist and deontological, as long as the matter is sufficiently concrete. Moreover, the available evidence, both from self-reports and neuroimaging data, suggests that people's deontological and retributivist punitive judgments are predominantly emotional, driven by feelings of anger or "outrage."

Emotion and the Moral Condemnation of Harmless Actions

According to consequentialists, actions are wrong because of their harmful consequences. In contrast, deontologists, along with many commonsense moralists, will condemn actions that do not cause harm in any ordinary sense. For example, a deontologist would likely say that it is wrong to break promises, regardless of whether doing so would have harmful consequences. Jonathan Haidt (Haidt, Koller, & Dias, 1993) has conducted a series of studies of moral judgments made in response to harmless actions. Two themes relevant to the present discussion emerge from this work. First, the moral condemnation of harmless action appears to be driven by emotion. Second, experience that encourages a more "cognitive" approach to moral decision making tends to make people less willing to condemn harmless actions.

Haidt and two Brazilian colleagues conducted a cross-cultural study of moral judgment using a large set of subjects varying in socioeconomic status (SES), nationality (Brazilian versus American), and age (children versus adults). The subjects were presented with a number of scenarios involving morally questionable, harmless actions:

1. A son promises his dying mother that he will visit her grave every week after she has died, but then doesn't because he is busy.
2. A woman uses an old American or Brazilian flag to clean the bathroom.
3. A family eats its dog after it has been killed accidentally by a car.
4. A brother and sister kiss on the lips.
5. A man masturbates using a dead chicken before cooking and eating it.

Subjects answered questions about each case: Is this action wrong? If so, why? Does this action hurt anyone? If you saw someone do this, would it bother you? Should someone who does this be stopped or punished? If doing this is the custom in some foreign country, is that custom wrong?

When people say that such actions are wrong, why do they say so? One hypothesis is that these actions are perceived as harmful, whether or not they really are (Turiel, Killen, & Helwig, 1987). Kissing siblings could cause themselves psychological damage. Masturbating with a chicken could spread disease, etc. If this hypothesis is correct, then we would expect people's answers to the question "Does this action hurt anyone?" to correlate with their degree of moral condemnation, as indexed by affirmative answers to the questions: "Is this wrong?" "Should this person be stopped or punished?" "Is it wrong if it's the local custom?" Alternatively, if emotions drive moral condemnation in these cases, then we would expect people's answers to the question "If you saw this, would it bother you?" to better predict their answers to the moral questions posed. As expected, Haidt and colleagues found that an affirmative answer to the "Would it bother you?" question was a better predictor of moral condemnation than an affirmative answer to the harm question.[8]

Equally interesting were the between-group differences. First, the high-SES subjects in Philadelphia and Brazil were far less condemning than their low-SES counterparts, so much so that the high-SES groups in Philadelphia and Brazil resembled each other more than they resembled their low-SES neighbors. Second, people from less "westernized"[9] cities tended to be more condemning. Third, children in both places tended to be more condemning than adults. In other words, education (SES), westernization, and growing up were associated with more consequentialist judgments in response to the scenarios used here. These three findings make sense in light of the model of moral judgment we have been developing, according to which intuitive emotional responses drive prepotent moral intuitions while "cognitive" control processes sometimes rein them in. Education is to a large extent the development of one's "cognitive" capacities, learning to think in ways that are abstract, effortful, and often either nonintuitive or counterintuitive. The westernization factor is closely related. While westerners may not be any more "cognitively" developed than members of other cultures, the western tradition takes what is, from an anthropological perspective, a peculiarly "cognitive" approach to morality. Westerners are more likely than members of other cultures to argue for and justify their moral beliefs and values in abstract terms (P. Rozin, personal communication, 2/23/05). Moreover, western culture tends to be more plural-

istic than other cultures, explicitly valuing multiple perspectives and an intellectual awareness that alternative perspectives exist. Finally, the capacity for "cognitive control" continues to develop through adolescence (V. A., Anderson, Anderson, Northam, Jacobs, & Catroppa, 2001; Paus, Zijdenbos, Worsley, Collins, Blumenthal, et al., 1999). Children, like adults, are very good at feeling emotions such as anger, sympathy, and disgust, but unlike adults they are not very good at controlling their behavior when experiencing such feelings (Steinburg & Scott, 2003). Thus, as before, there seems to be a link between "cognition" and consequentialist judgment.

In this study, the connection between a reluctance to condemn and consequentialism is fairly straightforward. Consequentialists do not condemn harmless actions.[10] The connection between the tendency to condemn harmless actions and deontology is, however, less straightforward and more questionable. It is not obvious, for example, that deontologists are any more likely than consequentialists to condemn flag desecration or eating the family dog. Similar doubts apply to the case of kissing siblings and the man who masturbates with a dead chicken, although it's worth noting that Kant argued that incest, masturbation, bestiality, and pretty much every other form of sexual experimentation are against the moral law (Kant, 1930; Kant, 1785/1994). The broken promise case, however, is "downtown deontology." Of course, not all deontologists would condemn someone for harmlessly breaking a promise to one's deceased mother, but anyone who would condemn such behavior (without appealing in some way to consequences) is exhibiting characteristically deontological behavior.[11] In light of this, it is worth examining this case more closely, and it turns out that this case fits the pattern for the intergroup differences quite well. Among high SES adults, the percentage of subjects in each city who said that this action should be stopped or punished ranged from 3% to 7%, while the percentage of low SES adults who said the same ranged from 20% (Philadelphia) to 57% (Recife, Brazil). Likewise, among high SES adults, the percentage who said that this behavior would be wrong even if it were the local custom ranged from 20% to 28%, while the corresponding percentages for low SES subjects ranged from 40% to 87%. The tendency to condemn this behavior also decreased with westernization, and within every group children were more willing to condemn it than adults. (If you want someone to visit your grave when you're dead, you can't beat poor children from Recife, Brazil. Ninety-seven percent endorsed punishing or stopping people who renege on grave-visiting promises, and 100% condemn cultures in which doing so is the custom.) Thus the argument made earlier connecting "cognition" and consequentialism applies

specifically to the case in which moral condemnation is most characteristically deontological. Haidt et al. did not provide data regarding the "Would it bother you?" question for this case specifically, but the fact that this case was not an exception to the general "cognitive" pattern (less condemnation in the presence of "cognition"-boosting factors) suggests that it is unlikely to be an exception to the general emotion-related pattern (condemnation correlated with negative emotions).

More powerful and direct evidence for the role of emotion in condemning harmless moral violations comes from two more recent studies. In the first of these, Thalia Wheatley and Jonathan Haidt (2005) gave hypnotizable individuals a posthypnotic suggestion to feel a pang of disgust upon reading the word "often" (and to forget that they received this suggestion). The other subjects (also hypnotizable individuals) were given the same treatment, except that they were sensitized to the word "take." The subjects were then presented with scenarios, some of which involved no harm. In one scenario, for example, second cousins have a sexual relationship in which they "*take/often* go on weekend trips to romantic hotels in the mountains." The subjects who received the matching posthypnotic suggestion (i.e., read the word to which they were hypnotically sensitized) judged this couple's actions to be more morally wrong than the other subjects.

In a second experiment, Wheatley and Haidt used a scenario in which the person described did nothing wrong at all. It was the case of a student council representative who "*often* picks" (or "tries to *take* up") topics of broad interest for discussion. Many subjects who received matching posthypnotic suggestions indicated that his behavior was somewhat wrong, and two subjects gave it high wrongness ratings. Subjects said things like: "It just seems like he's up to something," "It just seems so weird and disgusting," and, "I don't know [why it's wrong], it just is." Again, we see emotions driving people to nonconsequentialist conclusions.

In a more recent study, Simone Schnall, Jonathan Haidt, and Gerald Clore (2004) manipulated feelings of disgust, not with hypnosis, but by seating subjects at a disgusting desk while they filled out their questionnaires. (The desk was stained and sticky, located near an overflowing trashcan containing used pizza boxes and dirty-looking tissues, etc.) These subjects responded to a number of moral judgment scenarios, including variations on the dog-eating and masturbation scenarios mentioned earlier. Here, as before, the disgust manipulation made people more likely to condemn these actions, though only for subjects who were rated as highly sensitive to their own bodily states.

Two Patterns of Moral Judgment

The experiments conducted by Greene et al., Small and Loewenstein, Baron et al., Kahneman et al., Carlsmith et al., Sanfey et al., de Quervain et al., and Haidt et al. together provide multiple pieces of independent evidence that deontological patterns of moral judgment are driven by emotional responses while consequentialist judgments are driven by "cognitive" processes. Any one of the results and interpretations described here may be questioned, but the convergent evidence assembled here makes a decent case for the association between deontology and emotion, especially since there is, to my knowledge, no empirical evidence to the contrary. Of course, deontologists may regard themselves and their minds as exceptions to the statistically significant and multiply convergent psychological patterns identified in these studies, but in my opinion the burden is on them to demonstrate that they are psychologically exceptional in a way that preserves their self-conceptions.

Why should deontology and emotion go together? I believe the answer comes in two parts. First, moral emotion provides a natural solution to certain problems created by social life. Second, deontological philosophy provides a natural "cognitive" interpretation of moral emotion. Let us consider each of these claims in turn.

First, why moral emotions? In recent decades many plausible and complementary explanations have been put forth, and a general consensus seems to be emerging. The emotions most relevant to morality exist because they motivate behaviors that help individuals spread their genes *within a social context*. The theory of kin selection explains why individuals have a tendency to care about the welfare of those individuals to whom they are closely related (Hamilton, 1964). Because close relatives share a high proportion of their genes, one can spread one's own genes by helping close relatives spread theirs. The theory of reciprocal altruism explains the existence of a wider form of altruism: Genetically unrelated individuals can benefit from being nice to each other as long as they are capable of keeping track of who is willing to repay their kindness (Trivers, 1971). More recent evolutionary theories of altruism attempt to explain the evolution of "strong reciprocity," a broader tendency to reward cooperative behavior and punish uncooperative behavior, even in contexts in which the necessary conditions for kin selection (detectable genetic relationships) and reciprocal altruism (detectable cooperative dispositions) are not met (Bowles & Gintis, 2004; Fehr & Rockenbach, 2004; Gintis, 2000). These theories explain the widespread human tendency to engage in cooperative behaviors (e.g., helping others and speaking honestly) and to avoid uncooperative

behaviors (e.g., hurting others and lying), even when relatives and close associates are not involved. Moreover, these theories explain "altruistic punishment," people's willingness to punish antisocial behavior even when they cannot expect to benefit from doing so (Boyd, Gintis, Bowles, & Richerson, 2003; Fehr & Gachter, 2002; Fehr & Rockenbach, 2004). Other evolutionary theories make sense of other aspects of morality. For example, the incest taboo can be explained as a mechanism for avoiding birth defects, which are more likely to result from matings between close relatives (Lieberman, Tooby, & Cosmides, 2003). Finally, the emerging field of cultural evolution promises to explain how moral norms (and cultural practices more broadly) develop and spread (Richerson & Boyd, 2005).

Such evolutionary accounts of moral phenomena have received a great deal of attention in recent years (Pinker, 2002; Sober & Wilson, 1998; Wright, 1994), and therefore I will not elaborate upon them here. I will simply assume that the general thrust of these theories is correct: that our most basic moral dispositions are evolutionary adaptations that arose in response to the demands and opportunities created by social life. The pertinent question here concerns the psychological implementation of these dispositions. Why should our adaptive moral behavior be driven by moral emotions as opposed to something else, such as moral reasoning? The answer, I believe, is that emotions are very reliable, quick, and efficient responses to recurring situations, whereas reasoning is unreliable, slow, and inefficient in such contexts. (see Sober & Wilson (1998, chap. 10) on altruistic emotions versus hedonistic reasoning.)

Nature doesn't leave it to our powers of reasoning to figure out that ingesting fat and protein is conducive to our survival. Rather, it makes us hungry and gives us an intuitive sense that things like meat and fruit will satisfy our hunger. Nature doesn't leave it to us to figure out that fellow humans are more suitable mates than baboons. Instead, it endows us with a psychology that makes certain humans strike us as appealing sexual partners, and makes baboons seem frightfully unappealing in this regard. And, finally, Nature doesn't leave it to us to figure out that saving a drowning child is a good thing to do. Instead, it endows us with a powerful "moral sense" that compels us to engage in this sort of behavior (under the right circumstances). In short, when Nature needs to get a behavioral job done, it does it with intuition and emotion wherever it can. Thus, from an evolutionary point of view, it is no surprise that moral dispositions evolved, and it is no surprise that these dispositions are implemented emotionally.

Now, onto the second part of the explanation. Why should the existence of moral emotions give rise to the existence of deontological philosophy?

To answer this question, we must appeal to the well-documented fact that humans are, in general, irrepressible explainers and justifiers of their own behavior. Psychologists have repeatedly found that when people don't know why they're doing what they're doing, they just make up a plausible-sounding story (Haidt, 2001; Wilson, 2002).

Recall, for example, the pantyhose experiment described earlier. The subjects didn't know that they were drawn to items on the right side of the display, but when they were asked to explain themselves, they made up perfectly rational, alternative explanations for their preferences (Nisbett & Wilson, 1977). In a similar experiment, Nisbett and Wilson (1977) induced subjects to prefer the laundry detergent Tide by priming them with word pairs like "ocean-moon" in a preceding memory test. When subjects explained their preferences, they said things like "Tide is the best-known detergent," or "My mother uses Tide," or "I like the Tide box." In an early experiment by Maier (Maier, 1931; Nisbett & Wilson, 1977), subjects had to figure out a way to tie together two cords hanging from the ceiling, a challenging task since the cords were too far apart to be reached simultaneously. The solution was to tie a heavy object to one of the cords so that it could swing like a pendulum. The subject could then hold onto one cord while waiting for the other one to swing into reach. Maier was able to help his subjects solve this problem by giving them a subtle clue. As he was walking around the room he would casually put one of the cords in motion. The subjects who were aided by this clue, however, were unaware of its influence. Instead, they readily attributed their insights to a different, more conspicuous cue (Maier's twirling a weight on a cord), despite the fact that this cue was demonstrated to be useless in other versions of the experiment.

In a similar experiment Dutton and Aron (Dutton & Aron, 1974; Wilson, 2002) had male subjects cross a scary footbridge spanning a deep gorge, after which they were met by an attractive female experimenter. Control subjects rested on a bench before encountering the attractive experimenter. The subjects who had just braved the scary bridge, with their sweaty palms and hearts a'pounding, were more than twice as likely as the control subjects to call the experimenter later and ask her for a date. These individuals (many of them, at any rate) interpreted their increased physiological arousal as increased attraction to the woman they had met.

The tendency toward post hoc rationalization is often revealed in studies of people with unusual mental conditions. Patients with Korsakoff's amnesia and related memory disorders are prone to "confabulation." That is, they attempt to paper over their memory deficits by constructing elaborate stories about their personal histories, typically delivered with great

confidence and with no apparent awareness that they are making stuff up. For example, a confabulating patient seated near an air conditioner was asked if he knew where he was. He replied that he was in an air-conditioning plant. When it was pointed out that he was wearing pajamas, he said. "I keep them in my car and will soon change into my work clothes" (Stuss, Alexander, Lieberman, & Levine, 1978). Likewise, individuals acting under posthypnotic suggestion will sometimes explain away their behaviors in elaborately rational terms. In one case, a hypnotized subject was instructed to place a lampshade on another person's head upon perceiving an arbitrary cue. He did as instructed, but when he was asked to explain why he did what he did, he made no reference to the posthypnotic suggestion or the cue: "Well, I'll tell you. It sounds queer but it's just a little experiment in psychology. I've been reading on the psychology of humor and I thought I'd see how you folks reacted to a joke that was in very bad taste" (Estabrooks, 1943; Wilson, 2002).

Perhaps the most striking example of this kind of post hoc rationalization comes from studies of split-brain patients, people in whom there is no direct neuronal communication between the cerebral hemispheres. In one study, a patient's right hemisphere was shown a snow scene and instructed to select a matching picture. Using his left hand, the hand controlled by the right hemisphere, he selected a picture of a shovel. At the same time, the patient's left hemisphere, the hemisphere that is dominant for language, was shown a picture of a chicken claw. The patient was asked verbally why he chose the shovel with his left hand. He answered, "I saw a claw and picked a chicken, and you have to clean out the chicken shed with a shovel" (Gazzaniga & Le Doux, 1978; Wilson, 2002). Gazzaniga and LeDoux argue that these sorts of confabulations are not peculiar to spilt-brain patients, that this tendency was not created when these patients' intercerebral communication lines were cut. Rather, they argue, we are all confabulators of a sort. We respond to the conscious deliverances of our unconscious perceptual, mnemonic, and emotional processes by fashioning them into a rationally sensible narrative, and without any awareness that we are doing so. This widespread tendency for rationalization is only revealed in carefully controlled experiments in which the psychological inputs and behavioral outputs can be carefully monitored, or in studies of abnormal individuals who are forced to construct a plausible narrative out of meager raw material.

We are now ready to put two and two together. What should we expect from creatures who exhibit social and moral behavior that is driven largely by intuitive emotional responses and who are prone to rationalization of

their behaviors? The answer, I believe, is deontological moral philosophy. What happens when we contemplate pushing the large man off the footbridge? If I'm right, we have an intuitive emotional response that says "no!" This nay-saying voice can be overridden, of course, but as far as the voice itself is concerned, there is no room for negotiation. Whether or not we can ultimately justify pushing the man off the footbridge, it will always *feel* wrong. And what better way to express that feeling of non-negotiable absolute wrongness than via the most central of deontological concepts, the concept of a *right*: You can't push him to his death because that would be a violation of his *rights*. Likewise, you can't let that baby drown because you have a *duty* to save it.

Deontology, then, is a kind of moral confabulation. We have strong feelings that tell us in clear and uncertain terms that some things *simply cannot be done* and that other things *simply must be done*. But it is not obvious how to make sense of these feelings, and so we, with the help of some especially creative philosophers, make up a rationally appealing story: There are these things called "rights" which people have, and when someone has a right you can't do anything that would take it away. It doesn't matter if the guy on the footbridge is toward the end of his natural life, or if there are seven people on the tracks below instead of five. If the man has a right, then *the man has a right*. As John Rawls (Rawls, 1971, pp. 3–4) famously said, "Each person possesses an inviolability founded on justice that even the welfare of society as a whole cannot override" and, "In a just society the rights secured by justice are not subject to political bargaining or to the calculus of social interests." These are applause lines because they make emotional sense. Deontology, I believe, is a natural "cognitive" expression of our deepest moral emotions.

This hypothesis raises a further question. Why just deontology? Why not suppose that all moral philosophy, even all moral reasoning, is a rationalization of moral emotion? (This is the strong form of the view defended by Jonathan Haidt, 2001, whose argument is the model for the argument made here.)[12] The answer, I think, is that consequentialist moral judgment is not driven by emotion, or at least it is not driven by the sort of "alarm bell" emotion that drives deontological judgment. The evidence presented earlier supports this hypothesis, suggesting that consequentialist judgment is less emotional and more "cognitive," but it doesn't explain why this should be so. I argued earlier that there is a natural mapping between the content of deontological philosophy and the functional properties of alarmlike emotions. Likewise, I believe that there is a natural mapping between the content of consequentialist philosophy and the functional

properties of "cognitive" processes. Indeed, I believe that consequentialism is inherently "cognitive," that it couldn't be implemented any other way.

Consequentialism is, by its very nature, systematic and aggregative. It aims to take nearly everything into account, and grants that nearly everything is negotiable. All consequentialist decision making is a matter of balancing competing concerns, taking into account as much information as is practically feasible. Only in hypothetical examples in which "all else is equal" does consequentialism give clear answers. For real-life consequentialism, everything is a complex guessing game, and all judgments are revisable in light of additional details. There is no moral clarity in consequentialist moral thought, with its approximations and simplifying assumptions. It is fundamentally actuarial.

Recall the definition of "cognitive" proposed earlier "Cognitive" representations are inherently neutral representations, ones that, unlike emotional representations, do not automatically trigger particular behavioral responses or dispositions. Once again, the advantage of having such neutral representations is that they can be mixed and matched in a situation-specific way without pulling the agent in multiple behavioral directions at once, thus enabling highly flexible behavior. These are precisely the sorts of representations that a consequentialist needs in order to make a judgment based on aggregation, one that takes all of the relevant factors into account: "Is it okay to push the guy off the bridge if he's about to cure cancer?" "Is it okay to go out for sushi when the extra money could be used to promote health education in Africa?" And so on. Deontologists can dismiss these sorts of complicated, situation-specific questions, but consequentialists cannot, which is why, I argue, that consequentialism is inescapably "cognitive."

Some clarifications: First, I am not claiming that consequentialist judgment is emotionless. On the contrary, I am inclined to agree with Hume (1978) that all moral judgment must have some affective component, and suspect that the consequentialist weighing of harms and benefits is an emotional process. But, if I am right, two things distinguish this sort of process from those associated with deontology. First, this is, as I have said, a weighing process and not an "alarm" process. The sorts of emotions hypothesized to be involved here say, "Such-and-such matters this much. Factor it in." In contrast, the emotions hypothesized to drive deontological judgment are far less subtle. They are, as I have said, alarm signals that issue simple commands: "Don't do it!" or "Must do it!" While such com-

mands can be overridden, they are designed to dominate the decision rather than merely influence it.

Second, I am not claiming that deontological judgment cannot be "cognitive." Indeed, I believe that sometimes it is. (See below.) Rather, my hypothesis is that deontological judgment is affective at its core, while consequentialist judgment is inescapably "cognitive." One could, in principle, make a characteristically deontological judgment by thinking explicitly about the categorical imperative and whether the action in question is based on a maxim that could serve as a universal law. And if one were to do that, then the psychological process would be "cognitive." What I am proposing, however, is that this is not how characteristically deontological conclusions tend to be reached, and that instead they tend to be reached on the basis of emotional responses. This contrasts with consequentialist judgments which, according to my hypothesis, cannot be implemented in an intuitive, emotional way. The only way to reach a distinctively consequentialist judgment (i.e., one that doesn't coincide with a deontological judgment) is to actually go through the consequentialist, cost-benefit reasoning using one's "cognitive" faculties, the ones based in the dorsolateral prefrontal cortex (Greene et al., 2004).

This psychological account of consequentialism and deontology makes sense of certain aspects of their associated phenomenologies. I have often observed that consequentialism strikes students as appealing, even as tautologically true, when presented in the abstract, but that its appeal is easily undermined by specific counterexamples. (See the earlier discussion contrasting people's real-world motives and abstract justifications for punishment.) When a first-year ethics student asks, "But isn't it obvious that one should do whatever will produce the most good?" all you have to do is whip out the *footbridge* case and you have made your point. Whatever initial "cognitive" appeal consequentialist principles may have is quickly neutralized by a jolt of emotion, and the student is a newly converted deontologist: "Why is it wrong to push the man off the footbridge? Because he has a *right,* an inviolability founded on justice that even the welfare of society as a whole cannot override!" Then it's time for a new counterexample: "What if the trolley is headed for a detonator that will set off a nuclear bomb that will kill half a million people?" Suddenly the welfare of society as a whole starts to sound important again. "Cognition" strikes back with a more compelling utilitarian rationale, and the student is appropriately puzzled. As this familiar dialectic illustrates, the hypothesis that deontology is emotionally based explains the "*NEVER!*—except

sometimes" character of rights-based, deontological ethics. An alarmlike emotional response presents itself as unyielding and absolute, until an even more compelling emotional or "cognitive" rationale comes along to override it.

This hypothesis also makes sense of certain deontological anomalies, which I suspect will turn out to be the "exceptions that prove the rule." I have argued that deontology is driven by emotion, but I suspect this is not always the case. Consider, for example, Kant's infamous claim that it would be wrong to lie to a would-be murderer in order to protect a friend who has taken refuge in one's home (Kant, 1785/1983). Here, in a dramatic display of true intellectual integrity, Kant sticks to his theory and rejects the intuitive response. (He "bites the bullet," as philosophers say.) But what is interesting about this bit of Kantian ethics is that it's something of an embarrassment to contemporary Kantians, who are very keen to explain how Kant somehow misapplied his own theory in this case (Korsgaard, 1996a). Presumably the same goes for Kant's views of sexual morality (Kant, 1930, pp. 169–171; Kant, 1994). Modern academics are no longer so squeamish about lust, masturbation, and homosexuality, and so Kant's old-fashioned views on these topics have to be explained away, which is not difficult, since his arguments were never terribly compelling to begin with (see the epigraph). If you want to know which bits of Kant contemporary Kantians will reject, follow the emotions.

Normative Implications

Psychological "Is" and Moral "Ought"

The hypotheses advanced here concerning the respective psychological bases of consequentialism and deontology could certainly be wrong, but whether they are right or wrong cannot be determined from the armchair. Rather, it is an empirical matter. And although these hypotheses remain open to empirical challenge, I am from here on going to assume that they are correct in order to explore their broader philosophical implications. Since most moral philosophers do not regard their views as contingent upon the outcomes of particular debates in experimental psychology, this assumption should not be regarded as unduly restrictive.

Indeed, moral philosophers tend to steer clear of scientific controversies whenever possible on the grounds that scientific details are largely irrelevant to their enterprise: Science is about what is, while morality is about what ought to be, and never the twain shall meet (Hume, 1740/1978; Moore, 1903). Contrary to this received moral wisdom, I believe that

science does matter for ethics, not because one can derive moral truths from scientific truths, but because scientific information can challenge factual assumptions on which moral thinking implicitly depends. The key point of contact between moral philosophy and scientific moral psychology is moral intuition. Moral philosophers from Plato (Plato, 1987) on down have relied on their intuitive sense of right and wrong to guide them in their attempts to make sense of morality. The relevance of science then is that it can tell us how our moral intuitions work and where they come from. Once we understand our intuitions a bit better we may view them rather differently. This goes not only for moralists who rely explicitly on moral intuitions (Ross, 1930), but also for moralists who are unaware of the extent to which their moral judgments are shaped by intuition.

In recent years, several philosophers and scientists have questioned the reliability of moral intuitions and argued that understanding the psychology of moral intuition has normative implications (Baron, 1994; Greene, 2003; Horowitz, 1998; Sinnott-Armstrong, 2006; Unger, 1996). I will do the same, but in the following more specific way. I will argue that our understanding of moral psychology, as described here, casts doubt on deontology as a school of normative moral thought.

Rationalism, Rationalization, and Deontological Judgment

Your friend Alice goes on many dates, and after each one she reports back to you. When she extols the people she likes and complains about the ones she dislikes, she cites a great many factors. This one is brilliant. That one is self-absorbed. This one has a great sense of humor. That one is a dud. And so on. But then you notice something: All the people she likes are exceptionally tall. Closer inspection reveals that after scores of dates over several years, she has not given the thumb's up to anyone who is less than six-foot-four, and has not turned down anyone over this height. (You plug Alice's dating data into your statistics software and confirm that height is a near perfect predictor of Alice's preferences.) Suddenly it seems that Alice's judgment is not what you had believed, and certainly not what she believes. Alice, of course, believes that her romantic judgments are based on a variety of complicated factors. But, if the numbers are to be believed, she basically has a height fetish, and all of her talk about wit and charm and kindness is mere rationalization.

What this example illustrates is that it's possible to spot a rationalizer without picking apart the rationalizer's reasoning. Instead you need do only two things: First, you have to find a factor that predicts the

rationalizer's judgments. Second, you have to show that the factor that predicts the rationalizer's judgments is not plausibly related to the factors that according to the rationalizer are the bases for his or her judgments. Using this strategy, I believe that one can make a pretty good case against rationalist versions of deontology such as Kant's; i.e., the ones according to which characteristically deontological moral judgments are justified in terms of abstract theories of rights, duties, etc. The case against such theories is already implicit in the empirical material presented earlier, but it is worth spelling it out.

The bulk of this chapter has been devoted to satisfying the first of the two requirements I listed, i.e., to identifying a factor, namely emotional response, that predicts deontological judgment. Next, we must consider the nature of the relationship between this predictive factor and the factors that according to rationalist deontologists are the bases for their judgments. By definition, a rationalist cannot say that that some action is right or wrong *because* of the emotions we feel in response to it. Nevertheless, as an empirical matter of fact (we are assuming), there is a remarkable correspondence between what rationalist deontological theories tell us to do and what our emotions tell us to do. Thus, in light of these data, there are a series of coincidences for which various rationalist deontologists must account. For example, according to Judith Jarvis Thomson (1986, 1990) and Frances Kamm (1993, 1996) (both of whom count as rationalists for our purposes), there is a complicated, highly abstract theory of rights that explains why it is okay to sacrifice one life for five in the *trolley* case but not in the *footbridge* case, and it *just so happens* that we have a strong negative emotional response to the latter case but not to the former. Likewise, according to Colin McGinn (1999) and Frances Kamm (1999), there is a theory of duty that explains why we have an obligation to help Singer's drowning child but no comparable obligation to save starving children on the other side of the world, and it *just so happens* that we have strong emotional responses to the former individuals but not to the latter. According to Kant (2002) and many other legal theorists (Lacey, 1988; Ten, 1987), there is a complicated abstract theory of punishment that explains why we ought to punish people regardless of whether there are social benefits to be gained in doing so, and it *just so happens* that we have emotional responses that incline us to do exactly that. The categorical imperative prohibits masturbation because it involves using oneself as a means (Kant, 1994), and it *just so happens* that the categorical imperative's chief proponent finds masturbation really, really disgusting (see epigraph). And so on.

Kant, as a citizen of eighteenth-century Europe, has a ready explanation for these sorts of coincidences: God, in his infinite wisdom, endowed people with emotional dispositions designed to encourage them to behave in accordance with the moral law. Kant famously avoided invoking God in his philosophical arguments, but it's plausible to think that his faith prevented him, along with nearly everyone else of his day, from being puzzled by the order and harmony of the natural world, including its harmony with the moral law. Moreover, in light of his background assumptions, you can't really blame Kant for trying to rationalize his moral intuitions. His intuitions derive from his human nature ("the moral law within"; Kant, 1788/1993), and ultimately from God. God's a smart guy, Kant must have thought. He wouldn't give people moral intuitions *willy nilly*. Instead, we must have the intuitions we have for *good reasons*. And so Kant set out to discover those reasons, if not by force of reason, then by feat of imagination.

Present-day rationalist deontologists, as citizens of the twenty-first century, cannot depend on the notion that God gave us our moral emotions to encourage us to behave in accordance with the rationally discoverable deontological moral truth. Instead, they need some sort of naturalistically respectable explanation for the fact that the conclusions reached by rationalist deontologists, as opposed to those reached by consequentialists, appear to be driven by alarmlike emotional responses. And their explanation needs to compete with the alternative proposed here, namely that rationalist deontological theories are rationalizations for these emotional responses—an explanation that already has an advantage, given that so much human behavior appears to be intuitive (Bargh & Chartrand, 1999) and there is a well-documented tendency for people to rationalize their intuitive behavior (Haidt, 2001; Wilson, 2002).

What sort of explanation can rationalist deontologists give? They will have to say, first, that the correspondence between deontological judgment and emotional engagement is not a coincidence and, second, that our moral emotions somehow track the rationally discoverable deontological moral truth. They can't say that our emotional responses are the *basis* for the moral truth, however, because they are *rationalists*. So they are going to have to explain how some combination of biological and cultural evolution managed to give us emotional dispositions that correspond to an independent, rationally discoverable moral truth that is not based on emotion.

Those charged with this task immediately face another disadvantage, which is the chief point I wish to make here. There are good reasons to

think that our distinctively deontological moral intuitions (here, the ones that conflict with consequentialism) reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth.

Take, for example, the *trolley* and *footbridge* cases. I have argued that we draw an intuitive moral distinction between these two cases because the moral violation in the *footbridge* case is "up close and personal" while the moral violation in the *trolley* case is not. Moreover, I have argued that we respond more emotionally to moral violations that are "up close and personal" because those are the sorts of moral violations that existed in the environment in which we evolved. In other words, I have argued that we have a characteristically deontological intuition regarding the *footbridge* case because of a contingent, nonmoral feature of our evolutionary history. Moreover, I have argued that the same "up close and personal" hypothesis makes sense of the puzzling intuitions surrounding Peter Singer's aid cases and the identifiable-victim effect, thus adding to its explanatory power.

The key point is that this hypothesis is at odds with any hypothesis according to which our moral intuitions in response to these cases reflect deep, rationally discoverable moral truths. Of course, the hypothesis I have advanced could be wrong. But do rationalist deontologists want to count on it? And do they have any more plausible positive explanations to offer in its place?

A similar hypothesis can explain our inclinations toward retributive punishment. Consequentialists say that punishments should only be inflicted insofar as they are likely to produce good consequences (Bentham, 1789/1982). Deontologists such as Kant (Kant, 2002), along with most people (Baron et al., 1993; Baron & Ritov, 1993), are retributivists. They judge in favor of punishing wrongdoers as an end in itself, even when doing so is unlikely to promote good consequences in the future. Is this a moral insight on their part or just a by-product of our evolved psychology? The available evidence suggests the latter.

As discussed earlier, it appears that the emotions that drive us to punish wrongdoers evolved as an efficient mechanism for stabilizing cooperation, both between individuals (Trivers, 1971) and within larger groups (Bowles & Gintis, 2004; Boyd et al., 2003; Fehr & Gachter, 2002). In other words, according to these models, we are disposed to punish because of this disposition's biological consequences. Moreover, natural selection, in furnishing us with this disposition, had a "choice," so to speak. On the one hand, Nature could have given us a disposition to punish by giving us, first, an innate desire to secure the benefits of future cooperation and, second, some means by which to recognize that punishing noncooperators is often a

good way to achieve this end. In other words, Nature could have made us punishment consequentialists. Nature's other option was to give us a direct desire to punish noncooperators as an end in itself, even if in some cases punishing does no (biological) good. As noted earlier, Nature faces this sort of choice every time it generates a behavioral adaptation, and in pretty much every case, Nature takes the more direct approach. Psychologically speaking, we desire things like food, sex, and a comfortable place to rest because they are pleasant (and because their absence is unpleasant) and not because we believe they will enhance our biological fitness. The disposition toward punishment appears to be no exception to this general pattern. Psychologically speaking, we punish primarily because we find punishment satisfying (de Quervain et al., 2004) and find unpunished transgressions distinctly unsatisfying (Carlsmith et al., 2002; Kahneman et al., 1998; Sanfey et al., 2003).

In other words, the emotions that drive us to punish are *blunt biological instruments*. They evolved because they drive us to punish in ways that lead to (biologically) good consequences. But, *as a by-product of their simple and efficient design*, they also lead us to punish in situations in which no (biologically) good consequences can be expected. Thus, it seems that as an evolutionary matter of fact, we have a taste for retribution, not because wrongdoers truly deserve to be punished regardless of the costs and benefits, but because retributive dispositions are an efficient way of inducing behavior that allows individuals living in social groups to more effectively spread their genes.

Of course it's possible that there is a coincidence here. It could be that it's part of the rationally discoverable moral truth that people really do deserve to be punished as an end in itself. At the same time, it could *just so happen* that natural selection, in devising an efficient means for promoting biologically advantageous consequences, furnished us with emotionally based dispositions that lead us to this conclusion; but this seems unlikely. Rather, it seems that retributivist theories of punishment are just rationalizations for our retributivist feelings, and that these feelings only exist because of the morally irrelevant constraints placed on natural selection in designing creatures that behave in fitness-enhancing ways. In other words, the natural history of our retributivist dispositions makes it unlikely that they reflect any sort of deep moral truth.

I should emphasize that I am not claiming that consequentialist theories of punishment are correct because the tendency to punish evolved in order to produce good consequences. These "good consequences" need only be good from a biological point of view, and to assume that our ends must

coincide with the ends of natural selection would be to commit the naturalistic fallacy in its original form (Moore, 1903). At the same time, I wish to make it clear that I am not asserting that any tendency that we have as an evolutionary by-product is automatically wrong or misguided. I wouldn't claim, for example, that it is wrong to love one's adopted children (who do not share one's genes) or to use birth control simply because these behaviors thwart nature's "intentions." My claim at this point is simply that it is unlikely that inclinations that evolved as evolutionary by-products correspond to some independent, rationally discoverable moral truth. Instead, it is more parsimonious to suppose that when we feel the pull of retributivist theories of punishment, we are merely gravitating toward our evolved emotional inclinations and not toward some independent moral truth.[13]

What turn-of-the-millennium science is telling us is that human moral judgment is not a pristine rational enterprise—that our moral judgments are driven by a hodgepodge of emotional dispositions, which themselves were shaped by a hodgepodge of evolutionary forces, both biological and cultural. Because of this, it is exceedingly unlikely that there is any rationally coherent normative moral theory that can accommodate our moral intuitions. Moreover, anyone who claims to have such a theory, or even part of one, almost certainly does not. Instead, what that person probably has is a moral rationalization.

It seems then that we have somehow crossed the infamous "is" "ought" divide.[14] How did this happen? Didn't Hume (1978) and Moore (1903) warn us against trying to derive an "ought" from an "is?" How did we go from descriptive scientific theories concerning moral psychology to skepticism about a whole class of normative moral theories? The answer is that we did not, as Hume and Moore anticipated, attempt to *derive* an "ought" from an "is." That is, our method has been *inductive* rather than *deductive*. We have inferred on the basis of the available evidence that the phenomenon of rationalist deontological philosophy is best explained as a rationalization of evolved emotional intuition (Harman, 1977).

Missing the Deontological Point

I suspect that rationalist deontologists will remain unmoved by the arguments presented here. Instead, I suspect, they will insist that I have simply misunderstood what Kant and like-minded deontologists are all about. Deontology, they will say, isn't about this intuition or that intuition. It's not defined by its normative differences with consequentialism. Rather, deontology is about taking humanity seriously. Above all else, it's about

respect for persons. It's about treating others as fellow rational creatures rather than as mere objects, about acting for reasons that rational beings can share; and so on (Korsgaard, 1996a, 1996b).

This is, no doubt, how many deontologists see deontology. However, this insider's view, as I have suggested, may be misleading. The problem, more specifically, is that it defines deontology in terms of values that are not distinctively deontological, though they may appear to be from the inside. Consider the following analogy with religion. When one asks a religious person to explain the essence of his religion, one often gets an answer like this: "It's about love, really. It's about looking out for other people, looking beyond oneself. It's about community, being part of something larger than oneself." This sort of answer accurately captures the phenomenology of many people's religion, but it is nevertheless inadequate for distinguishing religion from other things. This is because many, if not most, nonreligious people aspire to love deeply, look out for other people, avoid self-absorption, have a sense of a community, and be connected to things larger than themselves. In other words, secular humanists and atheists can assent to most of what many religious people think religion is all about. From a secular humanist's point of view, in contrast, what is distinctive about religion is its commitment to the existence of supernatural entities as well as formal religious institutions and doctrines. And they are right. These things really do distinguish religious from nonreligious practices, although they may appear to be secondary to many people operating from within a religious point of view.

In the same way, I believe that most of the standard deontological/Kantian self-characterizations fail to distinguish deontology from other approaches to ethics. (See also Kagan, 1997, pp. 70–78, on the difficulty of defining deontology.) It seems to me that consequentialists, as much as anyone else, have respect for persons, are against treating people as mere objects, wish to act for reasons that rational creatures can share, etc. A consequentialist respects other persons and refrains from treating them as mere objects by counting every person's well-being in the decision-making process. Likewise, a consequentialist attempts to act according to reasons that rational creatures can share by acting according to principles that give equal weight to everyone's interests, i.e., that are impartial. This is not to say that consequentialists and deontologists do not differ. They do. It's just that the real differences may not be what deontologists often take them to be.

What, then, distinguishes deontology from other kinds of moral thought? A good strategy for answering this question is to start with concrete

disagreements between deontologists and others (such as consequentialists) and then work backward in search of deeper principles. This is what I have attempted to do with the *trolley* and *footbridge* cases and other instances in which deontologists and consequentialists disagree. If you ask a deontologically minded person why it is wrong to push someone in front of speeding trolley in order to save five others, you will get characteristically deontological answers. Some will be tautological: "Because it's murder!" Others will be more sophisticated: "The ends don't justify the means." "You have to respect people's rights." As we know, these answers don't really explain anything, because if you give the same people (on different occasions) the *trolley* case or the *loop* case (see earlier discussion), they will make the opposite judgment, even though their initial explanation concerning the *footbridge* case applies equally well to one or both of these cases. Talk about rights, respect for persons, and reasons we can share are natural attempts to explain, in "cognitive" terms, what we feel when we find ourselves having emotionally driven intuitions that are at odds with the cold calculus of consequentialism. Although these explanations are inevitably incomplete, there seems to be "something deeply right" about them because they give voice to powerful moral emotions. However, as with many religious people's accounts of what is essential to religion, they don't really explain what is distinctive about the philosophy in question.

In sum, if it seems that I have simply misunderstood what Kant and deontology are all about, it's because I am advancing an alternative hypothesis to the standard Kantian/deontological understanding of what Kant and deontology are all about. I am putting forth an empirical hypothesis about the hidden psychological essence of deontology, and it cannot be dismissed *a priori* for the same reason that tropical islanders cannot know *a priori* whether ice is a form of water.

Evolutionary Moral Psychology and Anthropocentric Morality

Earlier I made a case against rationalist deontology—the idea that our deontological moral intuitions can be justified by abstract theories of rights, duties, etc. There are, however, more modest forms of deontology. Rather than standing by our moral intuitions on the assumption that they can be justified by a rational theory, we might stand by them just because they are *ours*. That is, one might take an *anthropocentric* approach to morality (see Haidt & Bjorklund, volume 2), giving up on the Enlightenment dream of deriving moral truths from first principles and settling instead for a morality that is contingently human.

This is the direction in which moral philosophy has moved in recent decades. Virtue ethics defines moral goodness in terms of human character (Crisp and Slote, 1997; Hursthouse, 1999). Like-minded "sensibility theorists" regard being moral as a matter of having the right sort of distinctively human sensibility (McDowell, 1988; Wiggins, 1987). Ethicists with a more metaphysical bent speak of moral properties that are "response dependent" (Johnston, 1995), moral sentiments that correspond to "quasi-real" moral properties (Blackburn, 1993), and moral properties that are "homeostatic clusters" of natural properties (Boyd, 1988). Even within the Kantian tradition, many emphasize the "construction" of moral principles that rather than being true, are "reasonable for us" (Rawls, 1995), or, alternatively, the normative demands that follow from our distinctively human "practical identities" (Korsgaard, 1996b).

In short, moral philosophy these days is decidedly anthropocentric in the sense that very few philosophers are actively challenging anyone's moral intuitions. They acknowledge that our moral virtues, sensibilities, and identities may change over time, but they are not for the most part actively trying to change them.

The argument presented here makes trouble for people in search of rationalist theories that can explain and justify their emotionally driven deontological moral intuitions. But rationalist deontologists may not be the only ones who should think twice. The arguments presented here cast doubt on the moral intuitions in question regardless of whether one wishes to justify them in abstract theoretical terms. This is, once again, because these intuitions appear to have been shaped by morally irrelevant factors having to do with the constraints and circumstances of our evolutionary history. This is a problem for anyone who is inclined to stand by these intuitions, and that "anyone" includes nearly everyone.

I have referred to these intuitions and the judgments they underpin as "deontological," but perhaps it would be more accurate to call them nonconsequentialist (Baron, 1994). After all, you don't have to be a card-carrying deontologist to think that it's okay to eat in restaurants when people in the world are starving, that it's inherently good that criminals suffer for their crimes, and that it would be wrong to push the guy off the *footbridge*. These judgments are perfectly commonsensical, and it seems that the only people who are inclined to question them are card-carrying consequentialists.

Does that mean that all nonconsequentialists need to rethink at least some of their moral commitments? I humbly suggest that the answer is "yes". Let us consider, once more, Peter Singer's argument concerning the

moral obligations that come with affluence. Suppose, once again, that the evolutionary and psychological facts are exactly as I've said. That is, suppose that the *only* reason we say that it's wrong to abandon the drowning child but okay to ignore the needs of starving children overseas is that the former pushes our emotional buttons while the latter do not. And let us suppose further that the *only* reason that faraway children fail to push our emotional buttons is that we evolved in an environment in which it was impossible to interact with faraway individuals. Could we then stand by our commonsense intuitions? Can we, in good conscience, say, "I live a life of luxury while ignoring the desperate needs of people far away because I, through an accident of human evolution, am emotionally insensitive to their plight. Nevertheless, my failure to relieve their suffering, when I could easily do otherwise, is perfectly justified." I don't know about you, but I find this combination of assertions uncomfortable. This is not to say, of course, that I am comfortable with the idea of giving up most of my worldly possessions and privileges in order to help strangers. After all, I'm only human. But, for me at least, understanding the source of my moral intuitions shifts the balance, in this case as well as in other cases, in a more Singerian, consequentialist direction. As a result of understanding the psychological facts, I am less complacent about my all-too-human tendency to ignore distant suffering. Likewise, when I understand the roots of my retributive impulses, I am less likely to afford them moral authority. The same is true for whatever hang-ups I may have about deviant but harmless sexual behavior.

Taking these arguments seriously, however, threatens to put us on a second slippery slope (in addition to the one leading to altruistic destitution): How far can the empirical debunking of human moral nature go? If science tells me that I love my children more than other children only because they share my genes (Hamilton, 1964), should I feel uneasy about loving them extra? If science tells me that I am nice to other people only because a disposition to be nice ultimately helped my ancestors spread their genes (Trivers, 1971), should I stop being nice to people? If I care about myself only because I am biologically programmed to carry my genes into the future, should I stop caring about myself? It seems that one who is unwilling to act on human tendencies that have amoral evolutionary causes is ultimately unwilling to be human. Where does one draw the line between correcting the nearsightedness of human moral nature and obliterating it completely?

This, I believe, is among the most fundamental moral questions we face in an age of growing scientific self-knowledge, and I will not attempt to

address it here. Elsewhere I argue that consequentialist principles, while not true, provide the best available standard for public decision making and for determining which aspects of human nature it is reasonable to try to change and which ones we would be wise to leave alone (Greene, 2002; Greene & Cohen, 2004).

Notes

Many thanks to Walter Sinnott-Armstrong, Jonathan Haidt, Shaun Nichols, and Andrea Heberlein for very helpful comments on this chapter.

1. Kohlberg was certainly partial to deontology and would likely say that it is more "cognitive" than consequentialism.

2. It turns out that determining what makes a moral dilemma "personal" and "like the footbridge case" versus "impersonal" and "like the trolley case" is no simple matter, and in many ways reintroduces the complexities associated with traditional attempts to solve the trolley problem. For the purposes of this discussion, however, I am happy to leave the personal-impersonal distinction as an intuitive one, in keeping with the evolutionary account given earlier. For the purposes of designing the brain imaging experiment discussed later, however, my collaborators and I developed a more rigid set of criteria for distinguishing personal from impersonal moral violations (Greene et al., 2001). I no longer believe that these criteria are adequate. Improving these is a goal of ongoing research.

3. It is worth noting that no brain regions, including those implicated in emotion, exhibited the opposite effect. First, it's not clear that one would expect to see such a result since the hypothesis is that everyone experiences the intuitive emotional response, while only some individuals override it. Second, it is difficult to draw conclusions from negative neuroimaging results because current neuroimaging techniques, which track changes in blood flow, are relatively crude instruments for detecting patterns in neural function.

4. Of course, some aid organizations deliberately pair individual donors with individual recipients to make the experience more personal.

5. First, when I say that this behavior cannot be rationally defended, I do not mean that it is logically or metaphysically impossible for a rational person to behave this way. Someone could, for example, have a basic preference for helping determined victims and only determined victims. I am assuming, however, that none of the subjects in this experiment have such bizarre preferences and that therefore their behavior is irrational. Second, I am not claiming that the general psychological tendency that produces this behavior has no "rationale" or that it is not adaptive. Rather, I am simply claiming that this particular behavior is irrational in this case. Few, if any, of the participants in this study would knowingly choose to respond to

the experimental manipulation (determined versus undetermined victim) by giving more to the determined victim. In other words, this experimental effect would have been greatly diminished, if not completely eliminated, had this experiment employed a within-subject design instead of a between-subject design.

6. I am assuming that within the domain of punishment, "deontological" and "retributivist" are effectively interchangeable, even though they are conceptually distinct. (For example, one could favor punishment as an end in itself, but in unpredictable ways that defy all normative rules.) So far as I know, all well-developed alternatives to consequentialist theories of punishment are, in one way or another, retributivist. Moreover, retributivism is explicitly endorsed by many noteworthy deontologists, including Kant (2002).

7. Some complications arise in interpreting the results of these two studies of "outrage" and punishment. It is not clear whether the "outrage" scale used by Kahneman et al. elicits a subjective report of the subject's emotional state or a normative judgment concerning the defendant's behavior. A skeptic might say that the so-called "outrage" rating is really just a rating of the overall moral severity of the crime, which, not surprisingly, correlates with the extent to which people think it warrants punishment.

The Carlsmith et al. study addresses this worry (though not intentionally), and suggests that it may have some validity. The outrage measure used in the Carlsmith et al. study asks explicitly for a subjective report: "How morally outraged were you by this offense?" And, perhaps as a result of this change in tactic, the connection between "outrage" and punitive judgment is weakened from near perfect to fairly strong. Note also that in choosing a strong word like "outrage" in a study of fairly mild, hypothetical crimes, the experimenters may have set the bar too high for subjective reports, thus weakening their results.

8. This result, however, only held for the subgroups that did the majority of the condemning. The subjects who were most reluctant to condemn harmless violations (chiefly high-SES, educated westerners) found harm where others did not and cited that as a reason for condemnation, an effect that Haidt has documented elsewhere and which he has dubbed "moral dumbfounding" (Haidt, Bjorklund, & Murphy, 2000).

9. "Westernization" refers to "the degree to which each of three cities [Philadelphia and two Brazilian cities, Porto Alegre and Recife] has a cultural and symbolic life based on European traditions, including a democratic political structure and an industrialized economy" (Haidt, Koller, & Dias, 1993, 615). Philadelphia is more westernized than Porto Alegre, which is more westernized than Recife.

10. A consequentialist might favor a prohibition against a class of actions, some of which are not harmful, if the prohibition produces the best available consequences. Likewise, a consequentialist might *pretend* to condemn (or publicly condemn, while

privately refraining from condemning) an action if this behavior were deemed beneficial.

11. Subjects were asked to justify their answers, and typical justifications for condemning this action did not appeal to consequences, but rather simply stated that it's wrong to break a promise.

12. Haidt, however, believes that philosophers may be exceptional in that they actually do reason their way to moral conclusions (Kuhn, 1991).

13. That is, a truth independent of the details of human moral psychology and natural events that shaped it.

14. Most agree that the "is"-"ought" divide can be crossed when the "is" amounts to a constraint on what can be done, and is *a fortiori* a constraint on what "ought" to be done. For example, if it *is* the case that you are dead, then it is not the case that you *ought* to vote. The move from "is" to "ought" discussed later, however, is more substantive and correspondingly more controversial.

2.1 Moral Cognition and Computational Theory

John Mikhail

I

In his path-breaking work on the foundations of visual perception, David Marr distinguished three levels at which any information-processing task can be understood and emphasized the first of these:

> Although algorithms and mechanisms are empirically more accessible, it is the top level, the level of computational theory, which is critically important from an information-processing point of view. The reason for this is that the nature of the computations that underlie perception depends more upon the nature of the computational problems that have to be solved than upon the particular hardware in which their solutions are implemented. (Marr, 1982, p. 27.)

I begin with Marr to call attention to a notable weakness of Joshua Greene's wonderfully ambitious and provocative essay: its neglect of computational theory. A central problem moral cognition must solve is to recognize (i.e., compute representations of) the deontic status of human acts and omissions. How do people actually do this? What is the theory that explains their practice?

Greene claims that "emotional response . . . predicts deontological judgment" (Greene, p. 154), but his own explanation of a subset of the simplest and most extensively studied of these judgments—intuitions about trolley problems—in terms of a personal-impersonal distinction is neither complete nor descriptively adequate (Mikhail, 2002), as Greene now acknowledges in a revealing footnote. As I suggest later, a more plausible explanation of these intuitions suggests that the human brain contains a computationally complex "moral grammar" (e.g., Dwyer, 1999; Harman, 2000; Mikhail, 2000; Mikhail, Sorrentino, & Spelke, 1998) that is analogous in certain respects to the mental grammars operative in other domains, such as language, vision, music, and face recognition (Jackendoff, 1994). If this is

correct, then Greene's emphasis on emotion may be misplaced, and at least some of his arguments may need to be reformulated.

II

Consider the following variations on the trolley problem, which I designed to study the computations underlying moral judgments (Mikhail, 2000).

Bystander
Hank is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Hank sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Hank is standing next to a switch, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the men. Unfortunately, there is a man standing on the side track with his back turned. Hank can throw the switch, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Hank to throw the switch?

Footbridge
Ian is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ian sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Ian is standing next to a *heavy object*, which he can throw *onto the track in the path of the train*, thereby preventing it from killing the men. Unfortunately, *the heavy object* is a man, standing *next to Ian* with his back turned. Ian can throw the *man*, killing him; or he can refrain from doing this, letting the five die. Is it morally permissible for Ian to throw the *man*?

Consensual Contact
Luke is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Luke sees what has happened: the driver of the train saw *a man* walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the *man*. It is moving so fast that *he* will not be able to get off the track in time. Fortunately, Luke is standing next to *the man*, *whom* he can throw *off the track out of the path of the train*, thereby preventing it from killing the *man*. Unfortunately, *the man* is *frail and* standing with his back turned. Luke can throw the man, *injuring* him; or he can refrain from doing this, letting the *man* die. Is it morally permissible for Luke to throw the *man*?

Disproportional Death
Steve is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Steve sees what has happened: the driver of the train saw a man walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the man. It is moving so fast that he will not be able to get off the track in time. Fortunately, Steve is standing next to *a switch*, which he can throw, that will turn the train onto a side track, thereby preventing it from killing the *man*. Unfortunately, there are *five men* standing on the side track with their backs turned. Steve can throw the switch, killing the *five men*; or he can refrain from doing this, letting the *one man* die. Is it morally permissible for Steve to throw the *switch*?

As is well known, problems like these can be shown to trigger widely shared deontic intuitions among demographically diverse populations, including young children (Gazzaniga, 2005; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Hauser, Cushman, Young, & Mikhail, forthcoming; Mikhail, 2002; Mikhail et al., 1998; Petrinovich & O'Neill, 1996; Petrinovich, O'Neill, & Jorgensen, 1993; Waldmann, in press). Here I wish to draw attention to some of their theoretical implications.

III

It is clear that it is difficult if not impossible to construct a descriptively adequate theory of these intuitions—and others like them in a potentially infinite series—based exclusively on the information given (Mikhail, 2000). Although each of these intuitions is triggered by an identifiable stimulus, how the mind goes about interpreting these hypothetical fact patterns and assigning a deontic status to the acts they depict is not something revealed in any obvious way by the scenarios themselves. Instead, an intervening step must be postulated: a pattern of organization of some sort that is imposed on the stimulus by the mind itself. Thus a simple perceptual model, such as the one implicit in Haidt's (2001) influential account of moral judgment, is not adequate for explaining these intuitions.[1] Instead, as is the case with language perception (Chomsky, 1964), an adequate perceptual model must be more complex (figure 2.1.1).

The expanded perceptual model in figure 2.1.1 implies that, like grammaticality judgments, permissibility judgments do not necessarily depend only on the superficial properties of an action description, but also on how that action is mentally represented. In addition, it suggests that the problem of descriptive adequacy in the theory of moral cognition may be divided

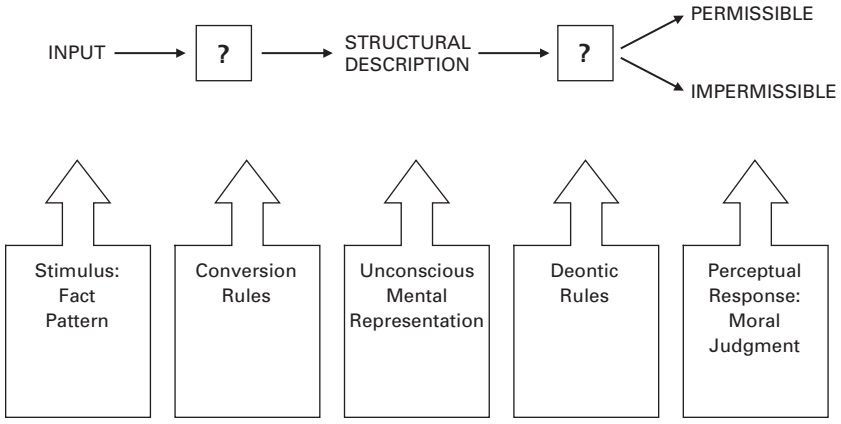

Figure 2.1.1
Expanded perceptual model for moral judgment (Mikhail, 2000).

into at least three parts: (1) the problem of describing the computational principles (deontic rules) operative in the exercise of moral judgment, (2) the problem of describing the unconscious mental representations (structural descriptions) over which those computational operations are defined, and (3) the problem of describing the chain of inferences (conversion rules) by which the stimulus is converted into an appropriate structural description.

IV

It is equally clear that Greene's own explanation of these intuitions is neither complete nor descriptively adequate. In a series of papers, Greene argues that people rely on three features to distinguish the bystander and footbridge problems: "whether the action in question (a) could reasonably be expected to lead to serious bodily harm, (b) to a particular person or a member or members of a particular group of people (c) where this harm is not the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107; see also Greene, 2005; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene & Haidt, 2002). Greene claims to predict trolley intuitions and patterns of brain activity on this basis. However, this explanation is incomplete because we are not told how people manage to interpret the stimulus in terms of these features; surprisingly, Greene leaves this crucial first step in the perceptual process (the step involving conversion rules) unanalyzed. In addition, Greene's account

is descriptively inadequate because it cannot explain even simple counterexamples like the consensual contact and disproportional death problems,[2]—let alone countless real-life examples that can be found in any casebook of torts or criminal law (Mikhail, 2002; Nichols & Mallon, 2006). Hence Greene has not shown that emotional response predicts these moral intuitions in any significant sense. Rather, his studies suggest that some perceived deontological violations are associated with strong emotional responses, something few would doubt or deny.

V

A better explanation of these intuitions is ready to hand, one that grows out of the computational approach Greene implicitly rejects. We need only assume people are "intuitive lawyers" (Haidt, 2001) and have a "natural readiness" (Rawls, 1971) to compute mental representations of human acts in legally cognizable terms. The footbridge and bystander problems, for example, can be explained by assuming that these problems trigger distinct mental representations whose relevant temporal, causal, moral, and intentional properties can be described in the form of a two-dimensional tree diagram, successive nodes of which bear a generation relation to one another that is asymmetric, irreflexive, and transitive (Goldman, 1970; Mikhail, 2000). As these diagrams reveal, the key structural difference between these problems is that the agent commits multiple counts of battery prior to and as a means of achieving his good end in the footbridge condition (figure 2.1.2), whereas in the bystander condition, these violations are subsequent and foreseen side effects (figure 2.1.3).

The computational or moral grammar hypothesis holds that when people encounter the footbridge and bystander problems, they spontaneously generate unconscious representations like those in figures 2.1.2 and 2.1.3. Note that in addition to explaining the relevant intuitions, this hypothesis has further testable implications. For example, we can investigate the structural properties of these representations by asking subjects to evaluate probative descriptions of these actions. Descriptions using the word "by" to connect individual nodes of the tree in the downward direction (e.g., "D turned the train by throwing the switch," "D killed the man by turning the train") will be deemed acceptable; by contrast, causal reversals using "by" to connect nodes in the upward direction ("D threw the switch by turning the train," "D turned the train by killing the man") will be deemed unacceptable. Likewise, descriptions using the phrase "in order to" to connect nodes in the upward direction along the vertical chain

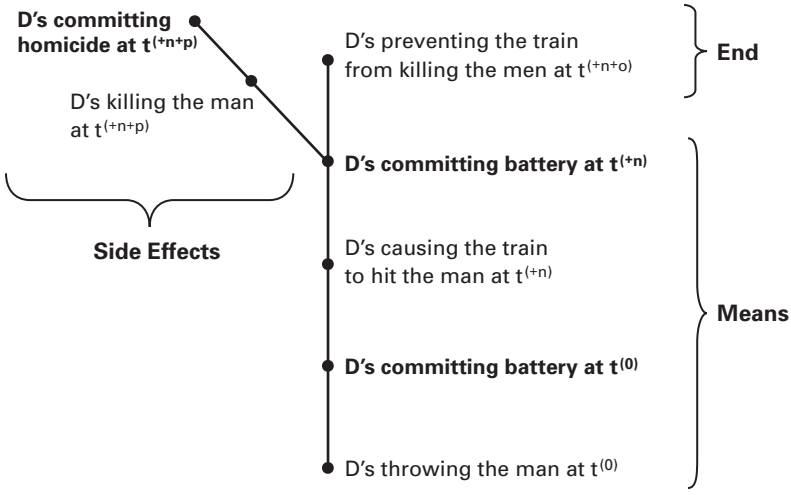

Figure 2.1.2
Mental representation of footbridge problem (Mikhail, in press).

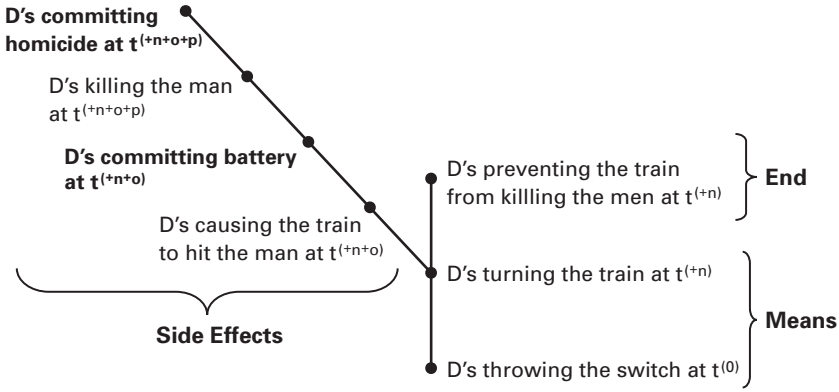

Figure 2.1.3
Mental representation of bystander problem (Mikhail, in press).

of means and ends ("D threw the switch in order to turn the train") will be deemed acceptable. By contrast, descriptions linking means with side effects ("D threw the switch in order to kill the man") will be deemed unacceptable. In short, there is an implicit geometry to these representations, which Greene and others (e.g., Sunstein, 2005) neglect, but which an adequate theory must account for (Mikhail, 2005).[3]

VI

The main theoretical problem raised by the computational hypothesis is how people manage to compute a full structural description of the relevant action that incorporates certain properties, such as ends, means, side effects, and prima facie wrongs like battery, when the stimulus contains no direct evidence for these properties. This is a poverty of the stimulus problem (Mikhail, 2006) that is similar in principle to determining how people manage to extract a three-dimensional representation from a two-dimensional stimulus in the theory of vision (e.g., Marr, 1982), or to determining how people recognize word boundaries in an undifferentiated auditory stimulus in the theory of language (e.g., Chomsky & Halle, 1968). Elsewhere I describe how these properties can be recovered from the stimulus by a sequence of operations that are largely mechanical (Mikhail, in press). These operations include (1) identifying the various action descriptions in the stimulus and placing them in an appropriate temporal and causal order, (2) applying certain moral and logical principles to their underlying semantic structures to generate representations of good and bad effects, (3) computing the intentional structure of the relevant acts and omissions by inferring (in the absence of conflicting evidence) that agents intend good effects and avoid bad ones, and (4) deriving representations of morally salient acts like battery and situating them in the correct location of one's act tree (Mikhail, 2000, 2002).[4] While each of these operations is relatively simple, the length, complexity, and abstract character of the process as a whole belies Greene's claim that deontological intuitions do not depend on "genuine" (p. 3 this volume), "complex" (p. 10 this volume), or "sophisticated abstract" (Greene & Haidt, 2002, p. 519) moral reasoning. In light of this and of Greene's failure to provide an adequate description of the computations that must be attributed to individuals to explain their moral intuitions, his reliance on characterizations like these seems unwarranted.

VII

Greene rejects the computational hypothesis largely on the strength of a single counterexample, namely, Thomson's (1986) ingenious loop case. "The consensus here," he says, "is that it is morally acceptable to turn the trolley . . . despite the fact that here, as in the footbridge case, a person will be used as a means" (Greene, this volume, ms p. 96; see also Greene et al., 2001, p. 2106). To test this assumption, I devised the following two scenarios (Mikhail, 2000) and discovered that no such consensus exists.

Loop 1 (Ned)
Ned is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Ned sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Ned is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. Unfortunately, the heavy object is a man, standing on the side track with his back turned. Ned can throw the switch, preventing the train from killing the men, but killing the man. Or he can refrain from doing this, letting the five die. Is it morally permissible for Ned to throw the switch?

Loop 2: (Oscar)
Oscar is taking his daily walk near the train tracks when he notices that the train that is approaching is out of control. Oscar sees what has happened: the driver of the train saw five men walking across the tracks and slammed on the brakes, but the brakes failed and the driver fainted. The train is now rushing toward the five men. It is moving so fast that they will not be able to get off the track in time. Fortunately, Oscar is standing next to a switch, which he can throw, that will temporarily turn the train onto a side track. There is a heavy object on the side track. If the train hits the object, the object will slow the train down, giving the men time to escape. Unfortunately, *there* is a man standing on the side track *in front of the heavy object* with his back turned. Oscar can throw the switch, preventing the train from killing the men, but killing the man; or he can refrain from doing this, letting the five die. Is it morally permissible for Oscar to throw the switch?

Unlike other trolley problems, on which roughly 85–95 percent of individuals agree, there is substantial disagreement over the permissibility of intervening in the two loop cases. For example, in the initial study utilizing these problems, only 48 percent of individuals judged Ned's throwing the switch to be permissible, whereas 62 percent judged Oscar's throwing the

switch to be permissible (Mikhail, 2002; see also Mikhail, 2000; Mikhail, Sorrentino, & Spelke, 1998). However, as these figures suggests, individuals did distinguish "Ned" and "Oscar" at statistically significant levels. These findings have since been replicated in a web-based experiment with several thousand subjects drawn from over 120 countries (Hauser, Cushman, Young, Jin, & Mikhail, 2007; see also Gazzaniga, 2005). Greene's account has difficulty explaining these findings, just as it has difficulty explaining the consensual contact and disproportionate death problems. All of these results, however, can be readily explained within a moral grammar framework (Mikhail, 2002).

VIII

In many respects, Greene's positive argument for an emotion-based approach to moral cognition has considerable plausibility. Nevertheless, some of the evidence he adduces in its favor appears to be weaker than he assumes. His reaction-time data, for instance, are inconclusive because the moral grammar framework makes the same predictions regarding people's reaction times and arguably provides a better explanation of them. One who permits throwing the man in the footbridge case must in effect overcome the prior recognition that this action constitutes an immediate and purposeful battery, and this process takes time; but one who prohibits throwing the switch in the bystander case need not override any such representation. Furthermore, while both the doing and forbearing of an action can be permissible without contradiction, the same is not true of the other two primary deontic operators (Mikhail, 2004). Hence Greene's reaction-time data can be explained by appealing to the cognitive dissonance resulting from the presence of a genuinely contradictory intuition in footbridge that is not present in bystander. By contrast, labeling the conflicting intuition a "prepotent negative emotional response" (Greene, this volume, ms. p. 103) does not seem explanatory, for reasons already discussed.

Some features of Greene's experimental design also may be questioned. For example, the fact that it takes longer to approve killing one's own child (crying baby case) than it does to condemn a teenage girl for killing hers (infanticide case) may not be entirely probative. Greene et al. (2001) appears to covary multiple parameters here (cost versus benefit and first person versus third person), undermining confidence in his results. More significantly, Greene does not appear to investigate considered judgments in Rawls' sense, that is, judgments "in which our moral capacities are most likely to be displayed without distortion" (Rawls, 1971, p. 47), in part

because most of his dilemmas are presented in the second person (e.g., "*You* are standing on a footbridge. . . . Is it appropriate for *you* to push the man?"). This presumably raises the emotional index of his scenarios and risks magnifying the role of exogenous factors.[5]

In addition, Greene does not appear to investigate deontic knowledge as such because he asks whether actions are appropriate instead of whether they are morally permissible.[6] That this question appears inapposite can be seen by considering the analogous inquiry in linguistics: asking whether an expression is "appropriate" rather than "grammatical." Chomsky (1957, p. 15) emphasized the importance of distinguishing grammatical from closely related but distinct notions like significant or meaningful, and the same logic applies here. Finally, whether one ought to perform a given action is distinct from whether the action is morally permissible, and Greene occasionally conflates this crucial distinction (see, e.g., Greene, Sommerville, Nystrom, Darley, & Cohen, 2001, p. 2105).

IX

These brief remarks are not meant to imply that Greene's project is without merit. On the contrary, I think his ideas are interesting, powerful, and at times even brilliant. His insight and creativity, clearly on display here, have helped give the field of moral psychology a much-needed boost. I would encourage him, however, to devote more effort to understanding the computational properties of moral cognition, in addition to its underlying mechanisms. Marr warned that "one has to exercise extreme caution in making inferences from neurophysiological findings about the algorithms and representations being used, particularly until one has a clear idea about what information needs to be represented and what processes need to be implemented" (Marr, 1982, p. 26). Without a better understanding of the rules and representations needed to explain widely shared moral intuitions, more caution would seem to be in order.

Notes

I wish to thank Josh Greene for writing such a thought-provoking essay and Walter Sinnott-Armstrong for his many helpful comments and suggestions on an earlier version of this commentary.

1. A notable feature of Haidt's social intuitionist model is that it provides no sustained analysis of the link between an eliciting situation and the intuitive response it generates (see Haidt, 2001, p. 814, figure 2).

2. Throwing the man in consensual contact is an action that "could reasonably be expected to lead to serious bodily harm to a particular person . . . where this harm is not the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107). In Greene's account, therefore, if I understand it correctly, this case should be assigned to his moral-personal category and judged impermissible. Yet in one experimental study, 93 percent of the participants found this action to be permissible (Mikhail, 2002). Conversely, while throwing the switch in the disproportional death problem is an action that "could reasonably be expected to lead to serious bodily harm to . . . a particular group of people," it is also "the result of deflecting an existing threat onto a different party" (Greene et al., 2001, p. 2107). In Greene's account, therefore, it should be assigned to his moral-impersonal category and judged permissible. Yet in the same study, 85 percent of the participants found this action to be impermissible. How do individuals manage to come to these conclusions? The answer cannot be the one proposed by Greene et al. (2001). However, it may be that I am misinterpreting the intended scope of Greene's personal-impersonal distinction, in which case clarification would be welcome.

3. Figures 2.1.2 and 2.1.3 also raise the possibility, which Greene does not consider, that deontic intuitions can be explained on broadly deontological (i.e. rule-based) grounds without reference to rights or duties. Put differently, they suggest that these concepts (and statements incorporating them, e.g., "Hank has a right to throw the switch," "Ian has a duty not to throw the man," "The man has a right not to be thrown by Ian"), while playing an important perspectival role in deontological systems, are conceptually derivative in a manner similar to that maintained by Bentham and other utilitarian theorists (Mikhail, 2000, 2004; Tuck, 1979).

4. In the footbridge problem, for example, one must infer that the agent must touch and move the man in order to throw him onto the track in the path of the train, and the man would not consent to being touched and moved in this manner because of his interest in self-preservation (and because no contrary evidence is given). By contrast, in the consensual contact problem, one naturally assumes that the man would consent to being thrown out of the way of the train, even though doing so will injure him. The computational hypothesis holds that when people respond intuitively to these problems, they do in effect make these inferences, albeit unconsciously.

5. Of course, if one wishes to study performance errors as such, then it may make sense to manipulate and enhance the influence of exogenous factors. This seems to be the approach adopted by Haidt and his colleagues (e.g., Wheatley & Haidt, 2005; Schnall, Haidt, & Clore, 2004) in the studies of theirs that Greene relies upon.

6. See Greene et al. (2001) 2105 Data Supplement—Supplemental Data at http://www.sciencemag.org/cgi/content/full/293/5537/2105/DC1 (last visited 9/25/2001).

2.2 Toward a Sentimentalist Deontology

Mark Timmons

Doubting Deontology: Greene's Challenge

In his splendid chapter, Joshua Greene launches an all-out assault on deontological philosophy (moral theory) from the vantage point of recent empirical work in psychology and brain science. As I see it, there are at least three main optio ns for the deontologist, ordered by how strongly they resist Greene's antideontology case. (1) Bold denial: deny that empirical work of the sort cited by Greene bears relevantly on normative moral theory generally and thus deontology in particular.[1] (2) Challenge: admit the philosophical relevance of empirical work, but challenge the empirical data brought forth by Greene by showing that there are flaws in the methodology, or that the results cannot be replicated, or something of this sort. (3) Acknowledgment: cautiously acknowledge the empirical data Greene presents, particularly his claim that commonsense deontological thinking is emotion laden, but explain how someone interested in developing a deontological moral theory can plausibly acknowledge the data in question.

I reject the first option on methodological grounds—I am sympathetic to Greene's claim that certain empirical findings are relevant to views in moral philosophy. As someone not trained in empirical science, I am not in a position to take up the challenge option and, anyway, I rather doubt that this option would pan out. So that leaves acknowledgment—the option I favor and which I will explore here, The main point I wish to make is that while Greene's arguments may work against some forms of deontology, they do not apply to other moral theories that are properly classified as deontological. Part of my strategy will be to refer to the work of some recent deontologists whose views do not seem to be affected by Greene's arguments, or at least whose views have the materials with which to mount a defense. I will proceed by considering what I take to be the

four main antideontology arguments Greene employs, which I will call the misunderstanding argument, the coincidence argument, the no normative explanation argument, and the sentimentalist argument.

Defining Deontology

Deontology covers a large and rather diverse range of normative moral theories and so there is no simple way of defining it.[2] Because a very common philosophical understanding of deontology is being challenged by Greene, I need to explain how I propose to understand this view. There are four elements that seem to capture the views of such deontologists as Kant, Prichard, and Ross, who ought to count as deontologists if anyone does.

Deontology is a normative theory about the relationship between right action and value (the good) according to which (1) some actions are wrong (contrary to one's duty) owing directly to such features of the action as that it is an instance of breaking a promise, killing an innocent person, lying, injustice, and so on.[3] Such features have a reason-providing authority that may explain the action's status as being morally required (a duty), morally wrong, or morally optional. (2) Humanity has a kind of value, the proper response to which is respect rather than "promotion"; respect for humanity will permit and sometimes require that in some circumstances we not promote the best consequences as gauged from the consequentialist perspective. So characterized, deontology is supposed to contrast with its main competitors, consequentialism and virtue ethics. To make the relevant contrasts clear, we may add two corollaries as part of our characterization. (3) The rightness of an action (its being a duty) is not in general constituted by facts about how much overall intrinsic value would (actually or probably) be brought about by the action. This distinguishes it from consequentialism. (4) Deontology also denies that the rightness of an action is always constituted by facts about the characters or motives of actual or ideal agents, as is characteristic of virtue ethics.

I am not sure that this characterization is broad enough in its positive claims to capture all deontological normative theories, but it will serve present purposes well enough.[4] Let us now turn to Greene's misunderstanding argument, which would challenge this standard characterization.

The Misunderstanding Argument

Throughout much of his chapter, Greene argues that characteristic deontological responses to a range of cases are emotion-laden, intuitive "gut

reactions" rather than the result of applying rules to cases or some other rational method of moral evaluation. By dwelling on these characteristic deontological intuitions, Greene attempts to build a case for the dual claim that (1) deontology is a philosophical attempt to rationalize a range of intuitive moral reactions—characteristically deontological reactions—and (2) the true essence of deontology is a certain pattern of psychological, intuitive (nonconsequentialist) responses to real and imagined cases calling for moral judgment. Thus deontologists have misunderstood the real essence of deontology.

Now one way to respond to Greene's appeal to intuitions is to point out that deontology need not attempt to capture all varieties of intuitive responses that he (Greene) is characterizing as deontological (a point which I'll develop in the next section) and that one can understand deontological views (at least in their Kantian versions) as having to do with respect for persons or what Kant called "humanity." This would be one way of characterizing the essence of deontology in terms of its distinctive content that would avoid Greene's deflationist psychological characterization of the view.[5] Greene anticipates this sort of move when toward the end of his chapter, he writes:

> Deontology, they [Kantians, especially of recent vintage] will say, isn't about this or that intuition. It's not defined by its normative differences with consequentialism. Rather, deontology is about taking humanity seriously. Above all else, it's about respecting persons. Its about treating others as fellow rational creatures rather than as mere objects, about acting for reasons that rational beings can share. (this volume, p. ••)

According to Greene, this response won't do because attempting to characterize deontology in terms of such values as humanity and the equal respect that is a fitting response to this value will fail to distinguish deontology from other moral theories, including consequentialism. Greene says:

> It seems to me that consequentialists, as much as anyone else, have respect for persons, are against treating people as mere objects, wish to act for reasons that rational creatures can share, etc. A consequentialist respects other persons and refrains from treating them as mere objects by counting every person's well-being in the decision-making process. Likewise, a consequentialist attempts to act according to reasons that rational creatures can share by acting according to principles that give equal weight to everyone's interests, i.e., that are impartial. (this volume, p. ••)

These remarks are supposed to block the attempt by philosophers to understand deontology in terms of a notion like respect for persons and thus

bolster the original two claims: (1) philosophers have not properly characterized a real difference between deontology and consequentialism and furthermore (2) if we dwell on the differences in commonsense moral reactions that differentiate deontological responses from consequentialist responses, we are led to real differences in the distinctive psychologies of these responses. Deontologists thus mischaracterize their view and so fail to understand its true nature. This is Greene's misunderstanding argument.

In response, a Kantian deontologist is going to insist that even if consequentialists can talk about the importance of equal respect for all persons and the value of humanity, there is still a nonconsequentialist, distinctively deontological conception of humanity that can serve as a basis (or at least a partial basis) for systematizing deontological moral principles of right conduct.[6] If this is right, then deontologists (at least those of a Kantian bent) can, after all, characterize their moral theory by appeal to the concept of humanity (properly interpreted). This point is, I think, well illustrated in the work of two recent deontologists, Robert Audi (2004) and T. M. Scanlon (1998). For brevity's sake, I'll make the point using Audi's view; later I make use of some ideas from Scanlon.

Audi's "Kantian intuitionism" involves an attempt to (1) systematize a Rossian theory of moral obligation (featuring a plurality of principles of prima facie duty) by deriving them from the humanity formulation of the categorical imperative and then (2) grounding the systematized set of principles in a theory of intrinsic value.[7] The overall result is a value-based deontology (Audi, 2004, p. 145). A value broad enough to serve as a ground is what, following Kant, Audi calls dignity, together with the attitude of respect that this status demands. So in the end, moral obligation is grounded in considerations of respect for persons[8]—a Kantian, nonconsequentialist notion of what constitutes respecting persons. Audi grants that the notions of human dignity and respect are, as he says "open-ended": "Their application is limited, however, in that they operate together and (so far as we are working within broadly Kantian constraints) both are fruitfully understood in reflective equilibrium with the categorical imperative, which in turn must be understood in reflective equilibrium with Rossian duties" (Audi 2004, p. 144; see also pp. 157–158).[9]

Thus Greene is right to claim that deontologists will likely insist that their theory is not about "this or that intuition." Rather, it is a distinctive normative theory that can be understood (at least for Kantians) as trying to capture in its theory of right conduct the appropriate way—a nonconsequentialist way—of respecting humanity. And isn't this sufficient for

properly characterizing deontology or at least Kantian strains of the view? So I don't think the deontologist—at least of a Kantian bent—is properly accused of not understanding the true essence of her theory. We can help strengthen this verdict by seeing how a deontologist can respond to some of Greene's other arguments.

Defending Deontology

In response to Greene's arguments against deontology, I think it is important to stress the importance of reflection in going from the raw materials of intuitive moral response to a moral theory. As Greene notes (in his remark about deontology not being about this or that intuition), the deontologist need not embrace unreflective moral reactions. Rather, the deontologist will claim that proper reflection on our various moral reactions of all sorts will yield a set of moral principles that we have reason to endorse, and that their overall structure will be deontological. Those principles in turn will help us discriminate "correct" or justified moral judgments from "incorrect" or unjustified moral judgments.

One clear example of the centrality of reflection in developing a version of deontology with a decidedly Kantian flavor is Scanlon's (1998) contractualist moral theory. The heart and soul of Scanlon's view is an account of proper moral thinking that is supposed to yield a set of moral principles (concerning our obligations to others) guided by the idea that a correct moral principle is one that we can justify to others. I won't elaborate the details of Scanlon's complex account of moral thinking here; I don't think they matter for present purposes. What does matter is that Scanlon proposes a roughly Kantian account of moral thinking that presumably has the power to move us beyond our unreflective moral judgments toward a deontological moral theory. Scanlon provides many examples of how his proposed methodology can take us beyond unreflective intuitions to a refined set of moral principles. I refer the reader in particular to what Scanlon says about the morality of taking human life (1998, pp. 199–200) and promising (1998, chap. 7). The centrality of reflection is also an important element of Audi's Kantian intuitionism.[10] Let us now turn to what are perhaps Greene's two strongest arguments against deontology.

The Coincidence Argument

Greene's critique is aimed at deontology per se, but in the latter part of his essay he begins referring to "rationalist deontology." Greene doesn't tell us what he means here by "rationalist", but I gather from what he says

in the section entitled, "Rationalism, Rationalization, and Deontological Judgment", that talk of rationalist deontology involves at least these two claims: (1) Moral judgments are beliefs and not emotions and (2) there is an independent realm of moral facts that serve as truth-makers for true moral beliefs. The gist of Greene's case against rationalist deontology seems to be the following. In light of the evidence for the claim that characteristically deontological judgments are highly emotion charged, together with the fact that there is a good evolutionary explanation for why we would have dispositions to make such judgments, the rationalist deontologist is going to have to explain how it is that there is this coincidence between our intuitive emotional reactions and an independent moral truth. Greene claims that Kant appeals to God in attempting to explain the coincidence and that any nontheological or metaphysical explanation is not likely to be forthcoming. For example, in connection with retributivist-deontological theories of punishment, Greene writes:

> [I]t seems that retributivist theories of punishment are just rationalizations for our retributivist feelings, and that these feelings only exist because of the morally irrelevant constraints placed on natural selection in designing creatures that behave in fitness-enhancing ways. In other words, the natural history of our retributivist dispositions makes it unlikely that they reflect any sort of deep moral truth. (this volume, pp. ••–••)

More generally about commonsense deontological thinking, he writes:

> My claim at this point is simply that it is unlikely that inclinations that evolved as evolutionary by-products correspond to some independent, rationally discoverable moral truth. (this volume, p. ••)

So what I am calling the coincidence challenge to the deontologist comes to this: explain why we should think that our intuitive moral responses with this sort of evolutionary history are tracking independent moral facts?

Here Greene stacks the deck against the deontologist by assuming that moral realism is essential to deontology. It isn't. Scanlon's constructivist version of deontology accepts an essentially cognitivist account of moral judgment but claims that moral truth is constituted by some set of idealized attitudes or responses.[11] In Scanlon's view in particular, and constructivist views in general, there are no independent moral facts to which true moral judgments must correspond. Rather, the constructivist takes moral truth to be a matter of ideally justified moral judgments and principles. If a Kantian deontologist can provide a plausible account of moral thinking that takes us from unreflective intuitive judgments to a set of moral

judgments which, ideally at least, constitute moral truth, then there would not seem to be some unexplained coincidence between the principles and judgments that a deontological theory yields and moral truth, so conceived. Thus, I don't see how the coincidence argument, as least as I understand it, is a challenge to deontology per se. (Maybe a realist-deontologist can mount a plausible defense against this objection, but I will leave that to others.)

Of course, Greene might say that there are more and less robust forms of independence and that his basic complaint applies mutatis mutandis to metaethical views that embrace the kind of modest independence featured in certain versions of moral constructivism like Scanlon's, according to which moral truth is independent of those moral principles we happen to accept. How would this version of the objection go? The idea would be that a constructivist deontologist owes us an explanation of why we have good normative reason to endorse deontological principles when the various commonsense deontological judgments seem to be the result of factors that are either morally irrelevant or at least not compatible with deontological reasons.

Again, I think this is a challenge that the deontologist can plausibly meet. For instance, Scanlon argues that we (nonamoralists) have reason to want our actions to be justifiable to others and that this reason provides a normative basis for explaining why moral reasons have the status and special importance they do seem to have.[12] So, in Scanlon's view, there is a good normative reason to endorse and care about the various moral reasons featured in a deontological theory.

The No Normative Explanation Argument

Another argument that Greene employs when he reviews the various bits of empirical evidence regarding commonsense deontological reactions is that attempts by deontologist philosophers to provide a normative characterization of the difference between deontological responses and consequentialist responses fail. In the trolley footbridge examples, for instance, one might propose (as a normative explanation of people's different reactions between these cases) the principle that one should not use another, innocent person merely as a means to some good end. Against this proposal, Greene cites Thomson's trolley loop example to show that this principle can't make the intuitively appropriate distinctions. Greene makes similar claims about other cases he discusses.

What does this apparent failure to come up with normative principles show about deontology? One might suppose that it only shows that

philosophers have not discovered the correct normative principles that would explain the different reactions in question. Perhaps they involve significant complexity and are difficult to formulate properly. The underlying principles that explain grammatical competence display a certain level of complexity that requires linguistic theory to uncover. Why not think that a similar thing is true regarding competent moral judgment and moral theory? Or, even if one is skeptical of there being such principles, why suppose that the deontologist is committed to holding that there must be principles of this sort? To sharpen these two points, let us consider some examples.

One might think of Scanlon's deontological view as compatible with the idea that there are underlying moral principles that have sufficient explanatory force of the sort Greene seems to demand. For Scanlon, moral principles "are general conclusions about the status of various kinds of reasons for action" (1998, p. 199) and arriving at justified moral principles is often a complex task requiring interpretation and judgment. In Scanlon's view, moral thinking about cases often involves refining overly simple moral generalizations such as "don't kill" and "don't lie" in which we appeal to a complex of relevant considerations bearing on a particular case under consideration in arriving at a more refined complex principle that we can see can be justified to others. Such principles may again be refined in light of the details of some further case. The process is thus one of refinement. This model of moral principles would presumably have us begin with fairly crude generalizations that prohibit intentional killing of innocent persons and refine them in light of the various morally relevant considerations featured in the trolley loop case, attempting to arrive at a refined moral principle that no one could reasonably reject. Again, Scanlon's view can arguably yield principles that explain the rightness or wrongness of an action by mentioning those morally relevant considerations that bear on the case at hand.

Audi's normative moral theory perhaps represents the most direct response to Greene's remarks about the limits of the Kantian requirement that we not treat others as mere means. According to Audi's Kantian intuitionism that I mentioned earlier, we are able to derive a plurality of Rossian duties from Kant's humanity formulation of the categorical imperative, and we can use the requirement that we not treat others as mere means as a guide for our deliberations in cases where these duties conflict. Now Greene might think that Audi's view is a sitting duck for cases like the trolley loop case, where it looks as if one ought (or is at least morally permitted) to use the lone innocent worker as a means for saving a greater number of innocent people. Why suppose that the principle that we are

to avoid treating people as mere means (the negative part of Kant's principle) states an exceptionless generalization? If one wants to claim that the lone innocent person is being used as a mere means (which is not entirely clear to me), then why not suppose that there can be difficult cases in which the all-things-considered morally correct thing to do is to treat the person as a mere means. In allowing for this possibility, we still have a very general moral principle that provides an important ceteris paribus constraint on morally permissible action and which can fulfill the unifying role envisioned by Audi.

Again, a deontologist might reject the idea that there are unrestricted moral principles that can be used to adjudicate conflicts among more particular principles of prima facie duty. This was Ross's view. He denied that there is a super principle that has relatively determinate implications and that can plausibly be used to adjudicate conflicts between prima facie duties. However, Ross's set of basic prima facie duties provides us with a moral framework within which we can reason about particular cases. In response to the trolley loop case, Ross would have us consider the details of the case in which the duties of beneficence and nonmaleficence seem most relevant and use practical judgment to adjudicate this particular conflict of prima facie duties.

Finally, it is worth mentioning that there are particularist versions of deontology of the sort we find in Prichard (1912/2002) that would deny that there are moral principles—of either the hard, exceptionless variety or of soft, ceteris paribus variety. So if Greene's objection here rests on the assumption that for the deontologist there must be moral principles that have the kind of determinacy sufficient to clearly and cleanly resolve hard moral cases, a particularist deontologist can simply deny the assumption.

The bottom line here is that Greene seems to suppose that the deontologist, in developing his or her theory, needs to come up with moral generalizations that (1) will distinguish deontological judgments (at least the ones a deontologist wants to endorse) from consequentialist judgments and (2) will articulate the normative basis for the judgments it implies. I think there are various ways in which this challenge can be met by deontologists who go in for moral principles, as illustrated by the views of Scanlon, Audi, and Ross. Also, a deontologist following Prichard need not embrace moral principles.

Developing Deontology

So far, I have been explaining how I think a rationalist-deontologist is likely to respond to Greene's arguments. In doing so, I have stressed (1)

the importance of a decidedly deontological conception of humanity as a normative basis for deontological principles of right conduct, (2) the role of reflection in going from intuitive moral responses to a normative moral theory, (3) constructivism about moral truth, and (4) the role that principles might (or might not) play in a deontological moral theory. I suppose that Greene might think that all of this maneuvering is not really going to help the deontologist at the end of the day because the entire basis of deontological thinking is emotion-laden intuitive responses. One remaining objection that is largely implicit in Greene's chapter but worth bringing out is the sentimentalist argument.

The Sentimentalist Argument

Deontology is committed to the idea that moral judgments are beliefs or are more cognitive than the evidence shows us; in short, deontology is committed to moral rationalism. However, in light of empirical evidence about people's intuitive moral judgments, a nonrationalist, sentimentalist account of them is more plausible than rationalist accounts. Thus, deontology is mistaken.

I am generally sympathetic to sentimentalism—as long as one doesn't overplay the role of sentiment in moral judgment (as I believe many sentimentalists are inclined to do)[13], and as long as one does not give up on the idea that moral judgments are a species of belief.[14] Although all of the versions of deontology that I know of have been embedded in a rationalist metaethic, I don't see why one cannot embrace sentimentalism (or expressivism) and go on to defend a deontological moral theory. Sentimentalism is a metaethical account about the nature of moral judgment; deontology is a normative theory about the right, the good, and their relation to one another. Although sentimentalism may seem to fit most comfortably with consequentialism, accepting the former metaethical view does not commit one to the latter normative moral theory.[15] So again, I don't see how (without further elaboration) the empirical facts about emotion-laden, intuitive moral reactions pose a threat to deontology. Indeed, I would suggest that the way to develop a deontological moral theory is to do so within the framework of a broadly sentimentalist metaethic.

Conclusion

There are other important and pressing challenges that Greene raises that I cannot take up here, including doubts about the evidential credentials of moral intuitions generally.[16] As far as I can tell, deontology per se is not

threatened by the empirical work cited by Greene; there are versions of deontology that can avoid Greene's arguments. The deontologist can appeal to a Kantian notion of respect for persons to systematize a set of soft ceteris paribus moral principles that are arrived at by an appropriate deontological method of moral deliberation that (if one accepts moral constructivism) constitute moral truth. Indeed, I suggest that empirical science can help the deontologist develop a sentimentalist metaethical framework to complement deontology. My colleague Michael Gill and I think the direction to go is toward a sentimentalist deontology, a view we plan to articulate and defend in the near future.

Notes

I wish to thank Robert Audi, Michael Gill, and especially Walter Sinnott-Armstrong for their many helpful comments on a previous version of this commentary.

1. Bold denial represents a denial of a "naturalizing" approach to philosophical problems.

2. See Freeman (2001) for a characterization of deontology that properly reflects this point.

3. With regard to metaphysical issues about the nature of right- and wrong-making properties or facts, one could distinguish among monist (Kant on some readings), pluralist (Ross 1930), and particularist (Prichard 2002) versions of deontology, but these intratheoretical differences will not be relevant for present purposes.

4. Points 1 and 2 may be too restrictive to count as requirements for a deontological theory, but all I am trying to do here is specify some common characteristics that we find in representatives of this kind of view, particularly those in the Kantian tradition who are the main target of Greene's criticisms.

5. Here I am taking Greene's misunderstanding argument as a challenge leveled against typical understandings of deontological moral theories in terms of their content. However, as Walter Sinnott-Armstrong pointed out to me, one might take Greene's challenge to be focused on the basis of deontological theories. I respond briefly to this latter form of challenge in a later section.

6. That the notion of equal respect is open ended and can be variously interpreted to fit with a variety of moral theories is a point nicely made by James Griffin (1986, p. 208; see also pp. 231, 239).

7. The kind of grounding in question is what Audi calls "ontic," which he contrasts with epistemic and inferential grounding (see Audi, 2004, p. 141).

8. Audi (2004, chap. 4, esp. pp. 141–145) holds that although moral obligations can be grounded in considerations of value, they do not need such ontic grounding to be known.

9. Audi (2004, p. 144) goes on to say that the relevant notion of dignity can be partially anchored in nonmoral notions of our rational capacities and sentience. Thus there are some nonmoral constraints on the interpretation of this concept.

10. Audi appeals to reflection in defending a conception of epistemological intuitionism in ethics (2004, pp. 45–48) and in explaining how Kant's humanity formulation can be used to derive principles of prima facie duty (2004, pp. 90–105).

11. Other Kantian constructivists include John Rawls (1971, 1980), Christine Korsgaard (1996b), and Onora O'Neill (1996).

12. See Scanlon (1998, chap. 3, esp. pp. 153–168).

13. For instance, I think it is a mistake for a sentimentalist to understand the content of ordinary moral judgments as being about certain sentiments as Gibbard does. See Nichols (2004b, chap. 4) for a critique of Gibbard's view and Gibbard's (2006) reply to Nichols.

14. According to the metaethical view that Terry Horgan and I favor (which we are calling "cognitivist expressivism"), moral judgments are genuine beliefs (hence we are cognitivists), but they are not descriptive beliefs (which puts us into the expressivist-sentimentalist camp). Our view is meant to challenge any kind of sharp reason versus sentiment dichotomy. See Horgan and Timmons (2006) and a forerunner of this view in Timmons (1999).

15. Blackburn (1993, p. 164) makes this point. I thank Michael Gill for this reference.

16. See Sinnott-Armstrong (2006) for objections to moral intuitionism based on findings in empirical psychology.

2.3 Reply to Mikhail and Timmons

Joshua D. Greene

Many thanks to John Mikhail and Mark Timmons for their thoughtful and challenging comments. Each of these authors teaches valuable lessons. The lessons they teach, however, are rather different, and so I will reply to them separately.

Reply to Mikhail

The first thing to note about John Mikhail's commentary on my chapter is that it is bold and incisive. Mikhail makes a number of strong claims about the limitations of my arguments, and many of these constitute serious challenges. The second thing to note about Mikhail's commentary on my chapter is that it is not really a commentary on my chapter. Rather, it is more or less a critique of my first neuroimaging paper (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), with some reference to subsequent interpretation (Greene & Haidt, 2002).

In my discussion here & advance a general empirical thesis: that deontological philosophy is largely a rationalization of emotional moral intuitions. While the results of my first neuroimaging study feature prominently in support of this thesis, my case is deliberately based on convergent evidence from many different experiments that bear directly on this issue (Baron, Gowda, & Kunreuther, 1993; Baron & Ritov, 1993; Carlsmith, Darley, & Robinson, 2002; de Quervain, Fischbacher, Treyer, Schellhammer, Schnyder, et al., 2004; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Haidt, Koller, & Dias, 1993; Kahneman, Schkade, & Sunstein, 1998; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Schnall, Haidt, & Clor, 2004; Small & Loewenstein, 2003, 2005; Wheatley & Haidt, 2005) as well as a number of experiments and theories that provide background support. In light of this, it is surprising that Mikhail focuses his attention exclusively on a single study. Despite this narrow focus, he raises a number

of important issues, and I will address them here. I will argue that our disagreements are real, but not as deep they may appear to be. I will make some concessions, stand my ground in other instances, and attempt to show the way forward.

According to Mikhail, a "notable weakness" of my research program is my "neglect of computational theory," which I "implicitly reject." I do not reject computational theories, implicitly or otherwise. (On the contrary, some of my *best friends* are computational theories, although I sometimes forget to call them on their birthdays.) It is true, however, that I have no computational theory to call my own. At least not yet. Mikhail, in contrast, does have a computational theory of moral judgment, which he briefly summarizes in his commentary. While my theory is "incomplete" and "descriptively inadequate," the relevant data can, according to Mikhail, be "readily explained within a moral grammar framework."

In my 2001 paper, my co-authors and I put forth a specific account of the standard trolley intuitions (that it's morally acceptable to trade one life for five in the *trolley* (bystander) case, but not in the *footbridge* case?). In his present commentary and elsewhere (Mikhail, 2000, forthcoming), Mikhail offers a competing account of this phenomenon. My current opinion is that both of these accounts are "incomplete" and "descriptively inadequate." Despite this, I think that both accounts reflect genuine insights. In the following pages I will attempt to explain what is right and not-so-right about our respective efforts to solve this problem.

The primary purpose of my first neuroimaging study was to find preliminary evidence for a set of related and fairly general ideas: Some moral decision making, I proposed, is driven by emotional intuitions, while other moral decision making is a product of abstract reasoning or is at least more "cognitive" (as in my chapter here). I proposed further that this duality in moral thought is reflected in the standard trolley intuitions. When a moral violation is "personal" (as in the *footbridge* case), it triggers a strong, negative emotional response that inclines people to judge against it. When a moral violation is "impersonal" (as in the *trolley* case), there is no comparable emotional response, and the judgment is made in a "cooler," more "cognitive" way. This general idea has an evolutionary rationale. I propose that we are disposed to respond emotionally to the kinds of personal moral violations that featured prominently in the course of human evolution, compared with moral violations that are peculiarly modern in ways that make them more "impersonal." I believe that these general ideas are correct and that they are supported by a substantial and growing body of data, including more recent data not covered in my chapter.

In designing our study, my collaborators and I were not committed to a specific hypothesis concerning what, exactly, triggers the hypothesized emotional responses to cases like the *footbridge* case. Nevertheless, our experimental design required that we take a stab at answering this question. (fMRI data are noisy, requiring repeated stimuli within each experimental condition. This meant that we had to define one class of dilemmas that are like the *footbridge* case and a distinct class of dilemmas that are like the *trolley* case. That meant that we had to say, if only provisionally, what the crucial difference between the *trolley* and *footbridge* cases is.) This is what we came up with: A moral violation is categorized as "personal" (as in the *footbridge* case) if it (1) could reasonably be expected to lead to serious bodily harm, (2) to a particular person or a member or members of a particular group of people (3) where this harm is not the result of deflecting an existing threat onto a different party. Moral violations that fail to meet these three criteria (as in the *trolley* case) are categorized as "impersonal." My co-authors and I suspected that this way of drawing the personal versus impersonal distinction would not fare well in the long run, and said so, describing it as a "first cut" and as "by no means definitive." It is now clear that this way of drawing this distinction does not work, but not necessarily for the reasons that Mikhail and others (Nichols & Mallon, 2006) have suggested, as I will explain.

First, this provisional hypothesis does not predict that all personal moral violations (as in the *footbridge* case) will be deemed inappropriate, or even deemed inappropriate by a majority of people. Rather, we claimed that personal moral violations trigger emotional responses that *incline* people to judge against them, but that these emotional responses can be overridden, particularly by utilitarian considerations. Thus, cases in which people judge "personal" moral violations to be appropriate, as in Mikhail's "consensual contact" case (pushing someone out of the way of an oncoming trolley) and Nichols and Mallon's "Catastrophe" case (killing one person to save billions of others) (Nichols & Mallon, 2006), pose no problem for our hypothesis. On the contrary, cases like these (i.e., cases in which there is an exceptionally strong utilitarian rationale for committing a personal moral violation) have provided essential reaction-time and neuroimaging data (Greene et al., 2004; Greene et al., 2001). Second, our provisional hypothesis does not predict that all impersonal moral violations will be deemed appropriate. Thus, for similar reasons, Mikhail's "disproportional death" case (turning a trolley onto five persons in order to save one) makes no trouble. According to our view, the absence of a personal moral violation means that there is little emotional response, which leads to a default,

utilitarian decision-making process. The judgment produced by this utilitarian process depends, of course, on the balance of costs and benefits and is not determined simply by the fact that the dilemma in question is "impersonal."

That said, there are several cases that *do* make trouble for our provisional hypothesis. The most damaging of these was, unbeknownst to us, already in the philosophical literature. This is Frances Kamm's "Lazy Susan" case (Kamm, 1996), which I will not discuss here. (I have since tested a version of this case and confirmed that it is indeed a counterexample.) Another case that makes trouble for our provisional hypothesis is Nichols and Mallon's teacups case (Nichols & Mallon, 2006). They presented subjects with modified versions of the *trolley* and *footbridge* cases in which teacups were substituted for people. Subjects took the action in the teacuppified *footbridge* case to be a more serious rule infraction than the action in the teacuppified *trolley* case, despite the fact that both of these cases are "impersonal" (because there is no bodily harm involved in either case). These results strongly suggest that there is at least some aspect of the *trolley-footbridge* effect that has nothing to do with personal violence. Finally, there are Mikhail's cases of *Ned* and *Oscar*. As Mikhail points out, the differences in people's responses to these two cases cannot be explained by appeal to any version of the personal/impersonal distinction.

Despite all this, the *general theory* presented in my 2001 paper (and elaborated upon in my chapter here) is well supported by published data, with more on the way. This general theory encompasses several claims:

1. Intuitive responses play an important role in moral judgment.
2. More specifically, intuitive responses drive people to give nonutilitarian responses to moral dilemmas that have previously been categorized as "personal."
3. This includes the *footbridge* case.
4. These intuitive responses are emotional (i.e., constituted or driven by emotions).
5. Cases like the *footbridge* case elicit negative emotional responses because they involve a kind of harm that is in *some sense* more personal than other kinds of harm.
6. We respond more emotionally to these "personal" harms because such harms, unlike others, were prominent during the course of human evolution.

Claim (6) remains a matter of evolutionary speculation. There is a great deal of evidence in favor of (1) and for the general importance of emotion

in moral judgment, much of it covered in my chapter here and elsewhere (Greene, 2005; Haidt, 2001). Regarding claims (2) and (4), there are the neuroimaging data from the 2001 paper itself. The "personal" cases produced increased activity in brain regions associated with emotion. These data have two principal limitations. First, these brain regions are not exclusively associated with emotion. Second, the activity observed in these brain regions could reflect incidental emotional activity that does not affect people's judgments. The reaction-time data presented alongside these neuroimaging data were collected in order to address this second concern.

The argument is as follows. If there is an emotional response that inclines people to say "no" to personal moral violations, then people should take longer when they end up saying "yes." If instead the emotional response is triggered later by the judgment itself, then it should have no effect on how long it takes people to make their judgments. And if the emotional response occurs in parallel with the judgment, then it could slow down people's judgments in a general way, but there is no reason to think that it would selectively interfere with one kind of answer. We found, as predicted, that "yes" answers are slower than "no" answers in response to personal moral dilemmas, with no comparable effect for impersonal dilemmas.

Mikhail claims that his theory can account for these data, but this is not so. According to my theory, people are slow to approve of personal moral violations because they must overcome a countervailing emotional response in order to do so. Mikhail suggests that they take longer because they "must overcome the prior recognition that this action constitutes an immediate and purposeful battery" (this volume, p. ••). While this could be true, it requires not only a major addition to Mikhail's theory, but an acknowledgment that there is real moral thinking (and not just noise and failures of "performance") outside of what Mikhail calls the "moral grammar." This is because Mikhail's theory makes no reference to any process that can overcome the initial deontic categorization produced by the moral grammar. Any such process is, from the point of view of Mikhail's theory, a *deus ex machina*. More specifically, Mikhail offers no positive explanation for why anyone would ever say that it's okay to push the guy off the footbridge. His theory, as stated, would be at its strongest if 100% of people said "no" to the *footbridge* case, which means that any "yes" answers given in response to this case are, as far as Mikhail's theory is concerned, just noise.

My view, in contrast, is that cases like the *footbridge* dilemma elicit competition between an intuitive emotional response and a more controlled

and "cognitive" utilitarian response, supported by activity in the dorsolateral prefrontal cortex (Greene et al., 2004). If Mikhail thinks that the output of the moral grammar is often forced to compete with some other kind of response, then his view is much closer to my "dual-process" view than it otherwise appears to be. The same is true if these competing responses are taken to be *part of* the moral grammar. However, if these competing responses are taken to be part of a specifically *moral* grammar, then Mikhail needs to explain why these processes bear such a striking neural resemblance to functionally similar control processes at work in nonmoral contexts (Greene et al., 2004). In other words, the mechanism behind the utilitarian judgments appear to be *domain general*. In either case, Mikhail's response to the reaction-time data seems to turn his theory into a special case of the general theory outlined here (claims 1–5). We agree that something about the action in the *footbridge* case triggers an intuitive response that inclines people to say "no," (claims 1–3) The only question then is whether we should call this intuitive response "emotional" (claim 4). In my chapter, I explain what I mean by "emotion." For a representation to be "emotional", it must be quick, automatic, etc., and also *valenced*. It must "say" that something is good or bad. And isn't that a perfect description of what Mikhail's moral grammar is supposed to deliver? A little voice that pops out of nowhere and says "No! That would be wrong!"

Well, we could spend an academic eternity arguing about whether the outputs of Mikhail's moral grammar should be called "emotional" by definition. However, that's not necessary because there are now three (and possibly four) new and independent pieces of evidence supporting my claim that emotional processes are responsible for generating the sorts of nonutilitarian responses we typically see in the *footbridge* case. More generally, these results (which are discussed in the following paragraphs) provide further evidence that there is a qualitative difference between the competing psychological processes that drive utilitarian versus nonutilitarian responses, making it even more of a strain to describe these processes as part of a single "grammar." (Of course, any cognitive system can be described as implementing a grammar, depending on what one means by "grammar.")

Patients with frontotemporal dementia (FTD) are known for their "emotional blunting" and lack of empathy. Recently, Mario Mendez and others presented FTD patients, Alzheimer's patients, and normal control subjects with versions of the *trolley* and *footbridge* dilemmas (Mendez, Anderson, & Shapira, 2005). A strong majority in all three groups said that they would

hit the switch in the *trolley* case, but, as predicted, the FTD patients diverged sharply in their responses to the *footbridge* case. While only 23 percent of the Alzheimer's patients and 19 percent of the normal control subjects said that they would push the guy off the *footbridge*, 57 percent of the FTD patients said they would do this, which is exactly what one would expect from patients who lack the emotional responses that drive ordinary people's responses to this case.

Michael Koenigs, Liane Young, and others have generated similar results in a recent unpublished study of patients with ventromedial prefrontal damage, another clinical population known for their emotional deficits (Damasio, 1994). They presented these patients with the set of "personal" moral dilemmas used in my 2001 study and found, as predicted, that these patients gave far more utilitarian answers than control patients and normal control subjects (Koenigs, Young, Cushman, Adolphs, Tranel, Damasio, & Hauser., forthcoming).

Valdesolo and DeSteno tested normal subjects with versions of the *trolley* and *footbridge* cases in conjunction with an emotion induction paradigm (Valdesolo & DeSteno, 2006). Subjects in the experimental condition watched a funny clip from *Saturday Night Live*. The control group watched a neutral film. Which film people watched had no significant effect on their responses to the *trolley* case, but the group that watched the SNL clip were about three times more likely to say that it's okay to push the man off the *footbridge*. Valdesolo and DeSteno predicted this for the following reason. If the negative response to the footbridge case is driven by a negative emotional response, then that response could be counteracted by a stimulus that produces a positive emotional response (i.e., a funny film clip).

Finally, My colleagues and I are currently conducting a cognitive load study using difficult personal moral dilemmas (like the *crying baby* case). So far, we are finding that burdening subjects with a cognitive load slows down their utilitarian moral judgments while it has no effect on (and possibly even speeds up) their deontological judgments. This is to be expected if deontological judgments, but not utilitarian judgments, are driven by intuitive emotional responses. (Whether this study provides additional evidence for the involvement of intuitive *emotional* processes depends, however, on whether one accepts my definition of "emotion.")

The neuroimaging and reaction time data presented in my 2001 paper strongly suggest that intuitive emotional responses incline people toward deontological responses to "personal" moral dilemmas (including the *footbridge* case). Nevertheless, this first study left ample room for doubt. The

studies just described, in contrast, leave little room for doubt. What, exactly, triggers these emotional responses, however, remains unknown. My original hypothesis is clearly wrong. Nevertheless, there is new evidence to suggest that the personal/impersonal distinction can be redrawn in a way that accounts for at least some of the data (claim 5). Fiery Cushman and colleagues have tested a version of the footbridge case in which the man on the footbridge, rather than getting pushed off the footbridge, can be dropped through a trapdoor operated by a nearby switch. They found that people judge saving the five to be more acceptable in the trapdoor version (Cushman, Young, & Hauser, forthcoming). (I independently ran the same experiment and got the same results.) These results strongly suggest that at least part of the *trolley-footbridge* effect has to do with "personalness," broadly construed.

What about Mikhail's alternative explanation of the *trolley-footbridge* effect? His theory can be understood on two levels. At the most general level, his theory is simply a descriptive restatement of the "doctrine of double effect," which turns crucially on the distinction between harming someone as a means and harming someone as a side effect. Mikhail's view goes further, however, in describing a plausible computational mechanism by which we might unconsciously distinguish means from side effect. [See also Michael Costa's similar theory (Costa, 1992).] Unfortunately, the means/side-effect distinction has severe limitations when it comes to explaining people's moral judgment behavior. And these limitations, of course, carry over to any more specific, computational account of how this distinction is applied. I will say, however, that Mikhail's theory is highly elegant and ingenious, and I suspect that there is something importantly right about it. Nevertheless, in its present form his theory doesn't work very well.

According to Mikhail's theory, subjects should say that any act of "intentional battery" (harming someone as a means) is wrong. However, as I have pointed out, people do not say this about the loop case (*Ned*), in which a person is used as a means to stop a trolley. In Mikhail's sample, about half of the subjects (48 percent) say that it is morally permissible for Ned to do this. While there may be no "consensus" in favor of running the guy over in this case, these data still make serious trouble for Mikhail's theory because approximately half the subjects *do the opposite of what his theory predicts*. To make matters worse, I have tested my own version of the *loop* case (using what I regard as less loaded language), and so far 73 percent of subjects say that it's morally acceptable to run the guy over. What's more, I have tested several other trolley variations that are structurally different

from the *loop* case, but that still involve killing someone as a means. In response to one of these cases, 84% of subjects (so far) say that it's okay to sacrifice the one person. As Mikhail points out, the means/side-effect distinction accounts for the fact that 62% of his subjects say that it's okay for *Oscar* (side-effect loop) to kill the one person while only 48% of his subjects say that it's okay for *Ned* (means loop) to kill the one person. And that's something. As I've said, the personal/impersonal distinction does nothing to explain this effect, and based on the data from these two cases, I'm inclined to believe that there is something right about Mikhail's theory. Nevertheless, explaining the 14% gap in people's responses to these two cases, however impressive, is a far cry from explaining the 60% to 80% gap between the *trolley* and *footbridge* cases. There is a lot more going on here.

Before closing, let me respond to a handful of Mikhail's remaining criticisms. First, Mikhail raises a worry about confounds in the design of my 2001 study. The particular one that he cites (killing one's own child versus someone else's killing her child) is not a concern because no dilemmas that differed along this dimension were contrasted in this study. There is, however, a more general worry about confounds in this study because the dilemmas respectively designated as "personal" and "impersonal" may differ in any number of unforeseen ways. We acknowledged this possibility in our paper, calling our personal/impersonal distinction a "first cut" and emphasizing the need for further research aimed at figuring out exactly what differences between these two sets of stimuli elicit the differences we observed in the fMRI and reaction-time data. In our subsequent work we have designed our experiments [e.g., the second analysis in our second neuroimaging paper (Greene et al., 2004) and the cognitive load study described earlier] to avoid such confounds, examining differences in neural activity and reaction time that are based on subjects' *responses* rather than the stimuli to which the subjects are responding.

Second, Mikhail raises concerns about our use of moral dilemmas presented in the second person, which may be more emotional than dilemmas presented in the third person. That may be so, but that's not a reason to ignore them in one's attempts to understand moral psychology. Moreover, the behavioral results generated using the second-person versions of the *trolley* and *footbridge* cases are broadly comparable to those generated using the third-person versions. It would be strikingly unparsimonious to suppose that the need one psychological theory to account for the *trolley-footbridge* effect in third-person cases and a completely different theory to account for the same effect in second-person cases. Finally, Mikhail raises a concern

about our asking subjects to judge whether actions are "appropriate" rather than "morally permissible." Because of this word choice, Mikhail claims, our study may not have been an investigation of "deontic knowledge as such." Our 2001 study used moral dilemmas that were identified as moral dilemmas by independent coders. Thus, whether or not our dilemmas required subjects to report their "deontic knowledge as such," we are confident that these dilemmas did require our subjects to make moral judgments, as ordinary people understand this activity.

Where to go from here? Based on data old and new, it is increasingly clear that intuitive emotional responses play a crucial role in the production of moral judgments, including those under consideration here. It is also increasingly clear that utilitarian considerations, supported by domain-general cognitive control mechanisms, can compete with, and in some cases override, these intuitive emotional responses. If these claims are correct, Mikhail's theory of "moral grammar" cannot serve as a general theory of moral judgment, or even as a general theory of trolley judgments. This is because his theory denies that emotions are anything other than sources of noise ("performance errors") and has no place for domain-general cognitive control mechanisms that can override intuitive responses.

Despite this, I believe that Mikhail's ideas are highly valuable. Within the framework of my "dual-process" model (Greene et al., 2004), there is an important unanswered question: What is the mechanism that triggers our intuitive emotional responses to cases like the *footbridge* case? My provisional hypothesis concerning the principles governing this mechanism is clearly wrong, although recent evidence suggests that the general idea behind this proposal ("personalness") has merit. It is equally clear that appeals to "personalness" will take us only so far. While the specific theory of "moral grammar" that Mikhail has offered has its limitations, I believe that the general ideas behind his theory have great merit and will prove useful in our attempts to understand the mechanisms behind our emotions.

Reply to Timmons

In his thoughtful and lucid commentary, Mark Timmons defends deontology with a twist. Why, he asks, can't we have an emotionally grounded deontology? This is an interesting proposal, worthy of serious consideration. Nevertheless, I remain skeptical.

Timmons identifies and evaluates four distinct arguments in my chapter, which he calls the *misunderstanding argument,* the *coincidence argument*, the

no normative explanation argument, and the *sentimentalist argument*. He also identifies a number of deontological moves that in his opinion can strengthen an enterprising deontologist's position. In what follows I will clarify and/or defend there arguments (which I regard as parts of a single argument). In the process I will explain why I believe the philosophical moves Timmons recommends are unlikely to help the deontological cause.

The conclusion of the *misunderstanding argument* is that the hidden essence of deontology is a psychological disposition toward emotionally driven moral judgments. This, as I understand it, is an empirical claim. Since Timmons has generously agreed to grant me my empirical claims, the question then becomes: What follows from this? If we grant that, psychologically speaking, intuitive emotional responses motivate deontological philosophy, does that mean that deontological philosophy is mistaken? Couldn't deontological thinking be the right kind of thinking, regardless of our psychological motives for embracing it? It could. In principle. Deontologists may someday construct or discover an elegant, self-justifying moral system that explains exactly how we are to value humanity and what sorts of things are right and wrong as a result. And they could claim further, as Kant did, that people's real-life psychological motives for judging and behaving rightly are irrelevant to moral theory. The fundamental principles of morals, they might argue, stand alone like mathematical theorems, independent of the messy world of psychology. Well, that is the deontological dream. But, as I have argued, keeping that particular dream alive requires one to posit a strange set of coincidences, which brings us to . . .

The *coincidence argument:* In response to the *coincidence argument*, Timmons makes two closely related moves: (1) he emphasizes the role of reflection in deontology and (2) takes a constructivist approach to moral truth. The dialectic goes like this. I say, "What are the chances that all these emotional responses are tracking the moral truth? Wouldn't that be a helluva coincidence?" To which Timmons replies, "Not at all. You're assuming some sort of hard-core realist metaethic, with the Moral Truth hovering above us in the Platonic ether. If our emotions were to track *that* kind of truth, that would be a bizarre coincidence indeed. But moral truth doesn't have to be that way. In a *constructivist* account of moral truth, the moral truth is just whatever comes out of a process of *rational reflection*. So it's no surprise if the output of that process (the moral truth) reflects the input to that process (our emotional intuitions)."

While this response sounds promising, it leaves the deontologist caught between two horns of a dilemma. But before we talk horns, we need to

distinguish between two types of deontology. "Ground-level" deontology, as I'll call it, is specifically committed to normative positions that are "characteristically deontological" and that are (ceteris paribus) at odds with consequentialism. Examples include Kant's infamous claim that it is wrong to lie to save someone's life (Kant, 1983) and the standard deontological view that it is wrong to push the guy off the *footbridge* in order to save five other people. It is this ground-level deontology that I had in mind in my chapter. There is, of course, a metaethical deontological tradition as well, which includes constructivist/contractualist philosophers like Rawls (1971) and Scanlon (1998). Their aim is to lay out a foundational theory of morality upon which a (nonutilitarian) ground-level theory can be "constructed." The construction process works as follows. We begin with our ordinary moral intuitions and commitments. Then we engage in some sort of rational reflection: "Are my current commitments consistent with rules that I would endorse if I were ignorant of my social position?" "Are my current commitments consistent with rules that no one could reasonably reject?" And through this reflective process our moral commitments are refined. At the ideal end of this reflective process, the moral principles to which we subscribe are the true ones.

So, here is the problem. Judgments based on emotional intuitions go into this reflective process. Do they come out? If they do come out, then we have what computer scientists call the GIGO problem: "garbage in, garbage out." (That's horn 1.) If they don't come out, then the output isn't necessarily deontological in the sense that matters (horn 2). Let us work through this argument using the now-familiar case of Peter Singer's utilitarian challenge to the affluent world (Singer, 1972). Since we are granting all of my empirical claims, let's assume, once again, that I'm completely right about the relevant psychology and its natural history: The *only* reason we are motivated to make a moral distinction between nearby drowning children and faraway starving children is that the former push our emotional buttons and the latter do not. And the *only* reason we exhibit this pattern of emotional response is because we did not evolve in an environment, like our current environment, in which we could have meaningful interactions with faraway strangers. Now, we take our characteristically deontological, emotion-based moral responses to these two cases (drowning child versus international aid) and feed them into the rational reflection process. If they somehow make it through, we have a problem. The so-called "moral truth" now reflects arbitrary features of our evolutionary history. GIGO. If, instead, our characteristically deontological intuitions do not survive this process of rational reflection, then in what sense is the

moral truth deontological? In the limiting case (which is consistent with the strong empirical assumptions I've been granted), all traces of our characteristically deontological intuitions are filtered out by the rational reflection process, and we are left with a ground-level utilitarian philosophy mounted upon a would-be deontological foundation. [This is more or less what John Harsanyi envisioned (Harsanyi, 1953, 1955).] My response to that is: Great! You can have your metaethical contractualism and constructivism as long as you are open to the possibility that the right ground-level theory is utilitarian and decidedly undeontological. As long as starving children get helped and people get shoved in front of speeding trolleys, that's all I care about.

Next we come to the *no normative explanation argument*. I point out that deontologists have a hard time justifying their judgments. Timmons points out that this does not mean that those judgments are necessarily wrong. It could be that just deontologists haven't yet worked out their arguments. Possible, sure. But I have also argued that these judgments can be explained in terms of patterns of emotional response and that these patterns reflect the influence of morally irrelevant factors. In light of this, wouldn't it be a strange coincidence if the correct moral theory just happened to map onto our moral emotions, which are sensitive to irrelevant factors? So this argument, too, brings us back to the *coincidence argument*.

Finally, we get to the *sentimentalist argument*. I have argued that deontologists who think they are rationalists are most likely rational*izers* of moral emotion. This is a problem for deontologists who insist on being genuine rationalists. But, Timmons, asks, why can't deontologists embrace the emotive foundations of their judgments? The answer, once again, is GIGO. Kant was opposed to emotion-based morality because emotions are fickle and contingent in oh-so-many ways (Kant, 1959). About that, he was right.