

# ETHICS AND INTUITIONS

Peter Singer<sup>1</sup>

## 1. Introduction

In one of his many fine essays, Jim Rachels criticized philosophers who “shoot from the hip.” As he put it:

The telephone rings, and a reporter rattles off a few “facts” about something somebody is supposed to have done. Ethical issues are involved – something alarming is said to have taken place – and so the “ethicist” is asked for a comment to be included in the next day’s story, which may be the first report the public will have seen about the events in question.

In these circumstances, Rachels noted, the reporters want a short pithy quote, preferably one that says that the events described are bad. The philosopher makes a snap judgment, and the result is something that reflects not “careful analysis” but “accepted wisdom.” Philosophers become “orthodoxy’s most sophisticated defenders, assuming that the existing social consensus must be right, and articulating its theoretical ‘justification’.” In contrast, Rachels argued, philosophers ought to “challenge the prevailing orthodoxy, calling into question the assumptions that people unthinkingly make.”<sup>2</sup>

Rachels’ own work in ethics lived up to that precept. To give just one of many possible examples, in what is probably his most cited article, on “Active and Passive Euthanasia”, he set out to criticize the common intuition that killing is worse than

---

<sup>1</sup> Ira W. DeCamp Professor of Bioethics. University Center for Human Values. Princeton University

<sup>2</sup> James Rachels, “When Philosophers Shoot from the Hip,” in Helga Kuhse and Peter Singer (eds.), *Bioethics: An Anthology* (Oxford: Blackwell, 1999), pp. 573, 575; reprinted from *Bioethics*, 5 (1991), pp. 67-71.

letting die. He showed that this distinction is influential in medicine, and is embodied in a statement from the American Medical Association. Then he convincingly argued that this is not an intuition on which we should rely.<sup>3</sup>

In both the papers I have mentioned, Rachels rejected the idea that the role of moral philosophers is to take our common moral intuitions as data, and seek to develop the theory that best fits those intuitions. On the contrary, he maintains, we should be ready to challenge the intuitions that first come to mind when we are asked about a moral issue. That is a view that I share, and one I have written about on several occasions over the years.<sup>4</sup> In the following pages I argue that recent research in neuroscience gives us new and powerful reasons for taking a critical stance toward common intuitions. But I will begin by placing this research in the context of our long search for the origins and nature of morality.

\*

In the Louvre Museum in Paris there is a black Babylonian column with a relief showing the sun god Shamash presenting the code of laws to Hammurabi. Such mythical accounts, bestowing a divine origin on morality, are common. In Plato's *Protagoras* there is an avowedly mythical account of how Zeus took pity on the hapless humans, who, living in small groups and with inadequate teeth, weak claws, and lack of speed, were no match for the other beasts. To make up for these deficiencies, Zeus gave humans a moral sense and the capacity for law and justice, so that they could live in larger communities and cooperate with one another. The biblical account of God giving the Ten Commandments to Moses on Mt. Sinai is, of course, another example.

---

<sup>3</sup> James Rachels, "Active and Passive Euthanasia," in Helga Kuhse and Peter Singer (eds.), *Bioethics: An Anthology* (Oxford: Blackwell, 1999), pp. 227-230; reprinted from *The New England Journal of Medicine*, 292 (1975), pp. 78-80.

<sup>4</sup> Starting with "Sidgwick and Reflective Equilibrium," *The Monist*, 58 (1974), pp. 490-517.

In addition to these mythical accounts, for at least 2500 years, and in different civilizations, philosophers have discussed and written about the nature of ethics. Plato himself was evidently not content with the account he offered in *Protagoras*, for in his dialogues he discusses several other possibilities. In the *Republic* alone, we have Thrasymachus's skeptical claim that the strong, acting in their own interests, impose morality on the weak, Glaucon's social contract model, and Socrates' proto-natural law defense of justice as the outcome of a harmony of the different parts of human nature.

Among the questions philosophers have considered are: whether ethics is objectively true, or relative to culture, or entirely subjective; whether human beings are naturally good; and whether ethics comes from nature or from culture. They have regarded such questions as having a practical, as well as theoretical, significance. Getting the answers right, they believe, will enable us to live in a better way.

Many of these thinkers were skilled observers of their fellow human beings, as well as being among the wisest people of their times. Consider, for example, the work of Mencius, Aristotle, Niccolo Machiavelli, Thomas Hobbes, and David Hume. There are many things about human nature that they understood very well. But none of them had the advantage of a modern scientific approach to these issues. Today we have that advantage. Hence it would seem odd if we could not improve on what they wrote.

In what follows, I summarize some of the new knowledge of ethics we now possess, knowledge that was not available to any of the great philosophers I have listed. Then I will consider what normative significance this new knowledge has. What, if anything, should it contribute to our debate over how we ought to act?

## 2. *Evolutionary Theory and the Origins of Morality*

The single most important advantage we have over the great moral philosophers of the past is our understanding of evolution and its application to ethics. Although the philosophers I have mentioned were able to free themselves from the myth of the divine origin of morality and to explain morality in naturalistic terms, they lacked a proper understanding of how our norms may have arisen by natural selection with the gene as the basic unit for the transmission of inherited characteristics between generations. Without this knowledge, they could observe our feelings and attitudes but not explain them adequately. To see what evolutionary theory can add to even the greatest of the pre-Darwinian thinkers who have speculated about the origins of morality, consider David Hume's discussion of morality in his justly celebrated *Treatise of Human Nature*.

Hume opens his discussion of justice by asking the question whether justice is a natural or an artificial virtue. In discussing that question he writes:

A man naturally loves his children better than his nephews, his nephews better than his cousins, his cousins better than strangers, where every thing else is equal. Hence arise our common measures of duty, in preferring the one to the other. Our sense of duty always follows the common and natural course of our passions.<sup>5</sup>

Hume gets very close to an evolutionary understanding of the common sense of duty, but he could not explain, as modern evolutionary theory can, why "the common and natural course of our passions" takes the form it does. We now understand that the genes that lead to the forms of love Hume describes are more likely to survive and spread among social mammals than genes that do not lead to preferences for one's relatives that are typically proportional to the proximity of the relationship. For we

---

<sup>5</sup> David Hume, *A Treatise of Human Nature*, ed. L.A. Selby-Bigge, Clarendon Press, Oxford, 1978, Bk. III, Pt. 2, Sec. i.

share more genes with our children than with our cousins, and more with our cousins than with strangers.

We can also now provide a deeper explanation of the truth of Hume's converse, and more controversial, observation that "there is no such passion in human minds as the love of mankind, merely as such, independent of personal qualities, of services, or of relation to ourself."<sup>6</sup> Much as we may regret it, most human beings lack a general feeling of benevolence for the strangers we pass in the street. In evolutionary terms, when we consider the species as a whole, the unit of selection is too large for natural selection to have much impact. Despite the picture books we had as children, early human life was not, by and large, a struggle for survival between humans and sabre-tooth tigers. It was much more often a struggle for survival between different human beings. There is no evolutionary advantage in concern for others simply because they are members of our species. In contrast to the selection of individual organisms within the species, which is going on all the time, selection between different species happens too slowly and too rarely to play much of a role in evolution.

Note, however, the factors that Hume lists as generating love for others: personal qualities, services, and relation to oneself. Relatedness we have already discussed. Personal qualities may generate positive feelings because they are likely to be of benefit to us, or to a small group to which we belong. In contrast to selection between species, which is rare and of little importance in evolution, selection within the species, between smaller, isolated breeding groups, happens much more often. These smaller groups do compete with each other and, in comparison with species, are relatively short-lived. The countervailing pressures of selection at the level of the individual or the gene would still apply, but less effectively. In some circumstances, there could be selective pressures that favor self-sacrifice for the benefit of the group. There would also, of course, be countervailing pressures favoring self-interested actions that do not benefit the group. If, however, the group develops a culture that rewards those who risk their own interests in order to benefit the group, and punishes

---

<sup>6</sup> Hume, *A Treatise of Human Nature*, Bk. III, Pt. 2, Sec. i.

those who do not, the cost-benefit ratio would be tilted so as to make benefiting the group more likely to be compatible with leaving offspring in the next generation.

The third exception that Hume mentioned was “services”. Here again he touches upon a focus of recent evolutionary theory, which has meshed with game theory in exploring such situations as the Prisoners’ Dilemma. This work enables us to give a fuller and more persuasive answer than Hume could to the question with which he began his discussion of justice.

Hume asked whether justice is a natural or an artificial virtue, and answered that it is an artificial one. By that he meant that “the sense of justice and injustice is not derived from nature, but arises artificially, though necessarily from education, and human conventions.” He adds that though the rules of justice are artificial, this does not mean that they are arbitrary. Justice is, for Hume, a human invention, though one that is “obvious and absolutely necessary.”<sup>7</sup> But justice is not, at least not in its origins, a human invention. We can find forms of it in our closer nonhuman relatives. A monkey will present its back to another monkey, who will pick out parasites; after a time the roles will be reversed. A monkey that fails to return the favor is likely to be attacked, or scorned in the future. Such reciprocity will pay off, in evolutionary terms, as long as the costs of helping are less than the benefits of being helped and as long as animals will not gain in the long run by “cheating”—that is to say, by receiving favors without returning them. It would seem that the best way to ensure that those who cheat do not prosper is for animals to be able to recognize cheats and refuse them the benefits of cooperation the next time around. This is only possible among intelligent animals living in small, stable groups over a long period of time. Evidence supports this conclusion: reciprocal behavior has been observed in birds and mammals, the clearest cases occurring among wolves, wild dogs, dolphins, monkeys, and apes.

---

<sup>7</sup> Hume, *A Treatise of Human Nature*, Bk. III, Pt. 2, Sec. i.

Many features of human morality could have grown out of simple reciprocal practices such as the mutual removal of parasites from awkward places. Suppose I want to have the lice in my hair picked out and I am willing in return to remove lice from someone else's hair. I must, however, choose my partner carefully. If I help everyone indiscriminately, I will find myself delousing others without getting my own lice removed. To avoid this, I must learn to distinguish between those who return favors and those who do not. In making this distinction, I am separating reciprocators and nonreciprocators and, in the process, developing crude notions of fairness and of cheating. I will strengthen my links with those who reciprocate, and bonds of friendship and loyalty, with a consequent sense of obligation to assist, will result.

This is not all. As we see with monkeys, reciprocators are likely to react in a hostile and angry way to those who do not reciprocate. More sophisticated reciprocators, able to think and use language, may regard reciprocity as good and "right" and cheating as bad and "wrong." From here it is a small step to concluding that the worst of the nonreciprocators should be driven out of society or else punished in some way, so that they will not take advantage of others again. Thus a system of punishment and a notion of desert constitute the other side of reciprocal altruism.

So Hume was not entirely wrong to say that justice is an artificial virtue, but he was not entirely right either. The basic rule of reciprocity, which includes the ability to detect cheats and the sense of indignation required to exclude them, is natural in the sense that it has evolved, is part of our biological nature, and is something we share with our closer nonhuman relatives. But the more detailed rules of justice typical of human, language-using societies are refinements on the instinctive sense of reciprocity, and so may be considered artificial.

Our biology does not prescribe the specific forms our morality takes. There are cultural variations in human morality, as even Herodotus knew.<sup>8</sup> Nevertheless, it seems likely that all these different forms are the outgrowth of behavior that exists in social animals, and is the result of the usual evolutionary processes of natural selection. Morality is a natural phenomenon. No myths are required to explain its existence.

### *3. How Humans Make Moral Judgments*

Against this background understanding of the origins of morality, I turn to some recent scientific research that helps us to understand more specific moral decisions and behavior. To explore the way in which people reach moral judgments, Jonathan Haidt, a psychologist at the University of Virginia, asked people to respond to the following story:

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

Haidt reports that most people are quick to say that what Julie and Mark did was wrong. They then try to give reasons for their answer. They may mention the dangers of inbreeding, but then recall Julie and Mark used two forms of birth control. Or they may suggest that the siblings could be hurt, even though it is clear from the

---

<sup>8</sup> See his account of the efforts of Darius, the Persian Emperor, to persuade people of different cultures to change their customs in respect of how to dispose of the dead, in *The Histories*, trans. Robin Waterfield, Oxford University Press, Oxford, 1998, Bk III, Ch. 38.

story that they were not. Eventually, many people say something like: “I don’t know, I can’t explain it, I just know it’s wrong.”<sup>9</sup> Evidently, it is the intuitive response that is responsible for the judgment these people reach, not the reasons they offer, for they stick to their immediate, intuitive judgment, even after they have withdrawn the reasons they initially offered for that judgment, and are unable to find better ones.

One example on its own would not show much, but Haidt has assembled an impressive body of evidence for the view that moral judgments in a variety of areas are typically the outcome of quick, almost automatic, intuitive responses. Where there is more deliberate, conscious reasoning, it tends to come after the intuitive response, and to be a rationalization of that response, rather than the basis for the moral judgment.<sup>10</sup>

If we turn to our growing knowledge of the parts of the brain involved in ethical decisions, we find a picture that is consistent with the conclusions that Haidt has drawn from studies of human behavior. Here we can begin with Antonio Damasio’s revealing discussion of the nineteenth century case of Phineas Gage. Gage was working on the American railroad when an explosion caused a 3-ft long iron rod to pass right through his brain. Astonishingly, Gage survived, and appeared to make a complete recovery, with no impairment to his reasoning or linguistic abilities. Yet it gradually became apparent that his character, previously steady and industrious, had changed. He became anti-social, and could not hold down a steady job as he had before.

---

<sup>9</sup> Jonathan Haidt, Fredrik Björklund, and Scott Murphy, “Moral Dumbfounding: When Intuition Finds No Reason,” unpublished manuscript (Department of Psychology, University of Virginia, 2000); and see further discussion in Jonathan Haidt, “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment,” *Psychological Review*, 108 (2001), pp. 814-834. I am indebted to Joshua Greene for drawing my attention to this, and other material discussed in this section, which draws on Chapter 3 of Greene’s Ph.D. thesis, *The Terrible, Horrible, No Good, Very Bad Truth about Morality, and What to Do About It*, unpublished manuscript (Department of Philosophy, Princeton University, 2002).

<sup>10</sup> Jonathan Haidt, “The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment,” *Psychological Review*, 108 (2001), pp. 814-834.

Gage's injury was to the ventromedial portion of the frontal lobes. More recent patients with damage to this area show the same combination of intact reasoning abilities but increased breaches of the usual moral and social standards. These patients appear to be emotionally deficient, not reacting in the usual way to gory scenes in which people's lives were lost or endangered. Damasio says of one of them that his predicament was "To know, but not to feel."<sup>11</sup> Brain-imaging studies have found a correlation between anti-social behavior and a deficiency in either the size of, or the amount of metabolic activity in, the prefrontal cortex.<sup>12</sup> In two patients where the damage to the ventromedial portion of the front lobes occurred early in life, the patients had much more marked psychopathic tendencies. They lied, stole and acted violently, and lacked any remorse.<sup>13</sup>

Further insight into the way in which we make moral judgments has come very recently from experiments using functional Magnetic Resonance Imaging, or fMRI, conducted by Joshua Greene and others at Princeton University. Greene designed the experiments to throw light on the way in which people respond to situations known in the philosophical literature as "trolley problems."<sup>14</sup> In the standard trolley problem, you are standing by a railroad track when you notice that a trolley, with no one aboard, is rolling down the track, heading for a group of five people. They will all be killed if the trolley continues on its present track. The only thing you can do to

---

<sup>11</sup> Antonio R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Grosset/Putnam, 1994), pp. 3-9, 34-51; the quote is on p. 45.

<sup>12</sup> Adrian Raine, Todd Lencz, Susan Bihrlé, Lori LaCasse, and Patrick Colletti, "Reduced Prefrontal Gray Matter Volume and Reduced Autonomic Activity in Antisocial Personality Disorder," *Archives of General Psychiatry*, 57 (2000), pp. 119-27; Adrian Raine, Monte S. Buchsbaum, Jill Stanley, Steven Lottenberg, Leonard Abel and Jacqueline Stoddard, "Selective Reductions in Prefrontal Glucose Metabolism in Murderers," *Biological Psychiatry*, 36 (1994), pp. 365-373.

<sup>13</sup> Steven W. Anderson, Antoine Bechara, Hanna Damasio, Daniel Tranel and Antonio R. Damasio "Impairment of Social and Moral Behavior Related to Early Damage in Human Prefrontal Cortex," *Nature Neuroscience*, 2 (1999), pp. 1032-1037.

<sup>14</sup> Phillipa Foot appears to have been the first philosopher to discuss these problems, in her paper "The Problem of Abortion and the Doctrine of the Double Effect," *Oxford Review*, 5 (1967), pp. 5-15; reprinted in James Rachels (ed.), *Moral Problems: A Collection of Philosophical Essays* (New York: Harper & Row, 1971), pp. 28-41. The classic article on the topic, however, is Judith Jarvis Thomson, "Killing, Letting Die, and the Trolley Problem." *The Monist* 59 (1976): 204-217..

prevent these five deaths is to throw a switch that will divert the trolley onto a side track, where it will kill only one person. When asked what you should do in these circumstances, most people say that you should divert the trolley onto the side track, thus saving four lives.

In another version of the problem, the trolley, as before, is about to kill five people. This time, however, you are not standing near the track, but on a footbridge above the track. You cannot divert the trolley. You consider jumping off the bridge, in front of the trolley, thus sacrificing yourself to save the imperilled people, but you realize that you are far too light to stop the trolley. Standing next to you, however, is a very large stranger. The only way you can stop the trolley killing five people is by pushing this large stranger off the footbridge, in front of the trolley. If you push the stranger off, he will be killed, but you will save the other five. When asked what you should do in these circumstances, most people say that you should not push the stranger off the bridge.

Many philosophers, including Judith Jarvis Thomson, see the problem posed by this pair of cases like this. In both cases you bring about the death of one person to save five, but we judge your action as right in the standard trolley case, and as wrong in the footbridge case. What is it that makes the difference between these two cases? These philosophers thus take the moral intuitions elicited by the cases as correct, and seek to justify them. But every time a seemingly plausible justifying principle has been suggested, other philosophers have produced variants on the original pair of cases that show that the suggested principle does not succeed in justifying our intuitive responses. For example, some philosophers suggested that the difference between the standard trolley case and the footbridge case is that in the latter the stranger is used as a means to save the others. Thus pushing the stranger off the footbridge violates the Kantian injunction not to use another person merely as a means, while throwing the switch does not. Unfortunately for proponents of this neat explanation, we can imagine a case in which throwing the switch does not cause the trolley to run down an altogether different track, but makes it go around a loop before

it reaches the five people threatened by it. On that loop, the very large stranger is lying. Because he is so large, his body will bring the trolley to a stop, but not before it kills him. To divert the trolley around this loop does use the stranger as a means to saving the life of the other five, but most people consider it would be right to do it. They thus judge this case as closer to the standard case of throwing the switch than to the case of pushing the stranger off the footbridge.

Unlike the many philosophers who have tried to justify our intuitions in these situations, Greene was more concerned to understand why we have them. He thought that the roots of the differing judgments we make about the two situations may lie in our different emotional responses to the idea of causing a stranger's death by throwing a switch on a railway track, and pushing someone to his or her death with our bare hands. As Greene puts it:

Because people have a robust, negative emotional response to the personal violation proposed in the footbridge case they immediately say that it's wrong... At the same time, people fail to have a strong negative emotional response to the relatively impersonal violation proposed in the original trolley case, and therefore revert to the most obvious moral principle, "minimize harm," which in turn leads them to say that the action in the original case is permissible.<sup>15</sup>

Greene used fMRI imaging, which provides a real-time image of activity in different parts of the brain, to test this hypothesis. He predicted that people asked to make a moral judgment about "personal" violations like pushing the stranger off the footbridge would show increased activity in areas of the brain associated with the emotions, when compared with people asked to make judgments about relatively "impersonal" violations like throwing a switch. But he also made a more specific prediction: that the minority of subjects who do consider that it would be right to push

---

<sup>15</sup> Greene, *The Terrible, Horrible, No Good, Very Bad Truth about Morality, and What to Do About It*, p. 178.

the stranger off the footbridge would, unless they were psychopaths, be giving this response in spite of their emotions, and therefore they would take longer to reach this judgment than those who say that it would be wrong to push the stranger off the footbridge, and also longer than they would take to reach a judgment in a case that did not arouse such strong emotional responses.

Greene's predictions were confirmed. When people were asked to make judgments in the "personal" cases, parts of their brains associated with emotional activity were more active than when they were asked to make judgments in "impersonal" cases. More significantly, those who came to the conclusion that it would be right to act in ways that involve a personal violation, but minimize harm overall – for example, those who say that it would be right to push the stranger off the footbridge – took longer to form their judgment than those who said it would be wrong to do so.<sup>16</sup>

When Greene looked more closely at the brain activity of these subjects who say "yes" to personal violations that minimize overall harm, he found that they show more activity in parts of the brain associated with cognitive activity than those who say "no" to such actions.<sup>17</sup> These are preliminary results, based on a limited amount of data. But let us assume that they are sound, and speculate on what might follow from them, in conjunction with the other scientific information relevant to the origins of ethics, as outlined above.

#### 4. Normative Implications

---

<sup>16</sup> Joshua D. Greene, R. Brian Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen, "An fMRI Investigation of Emotional Engagement in Moral Judgment," *Science*, 293 (2001), pp. 2105-2108. To be more specific: in personal moral dilemmas, the medial frontal cortex, posterior cingulate cortex, and angular gyrus/superior temporal sulcus are active. In impersonal moral dilemmas there is increased activity in the dorsolateral prefrontal cortex and parietal lobe.

<sup>17</sup> Joshua Greene and Jonathan Haidt, "How (and Where) Does Moral Judgment Work?," *Trends in Cognitive Sciences*, 6 (2002), pp. 517-523, and personal communications. To be more specific, those who accept the personal violation show more anterior dorsolateral prefrontal activity while those who reject it have more activity in the precuneus area.

Shortly after *The Origin of Species* appeared, Darwin wrote to a friend: “I have received in a Manchester newspaper rather a good squib, showing that I have proved ‘might is right’...”<sup>18</sup> Darwin knew, of course, that he had done nothing of the sort. The Social Darwinists committed the same fallacy when they argued against state interference with the free market on the grounds that protecting the poor and weak was interfering with natural selection. Assuming that we can define the term “natural” in a way that makes it meaningful to say that protecting the poor and weak interferes with natural selection, we would still need an ethical argument to say that it is wrong to do so. The direction of evolution neither follows, nor has any necessary connection with, the path of moral progress. “More evolved” does not mean “better”. No matter how often the fallacy of reading a moral direction into evolution has been pointed out, people still commit it, and it is not difficult to find otherwise excellent contemporary writers in evolutionary theory who continue to make this mistake.<sup>19</sup> Nevertheless, it is a mistake. So while I have claimed that evolutionary theory explains much of common morality, including the central role of duties to our kin, and of duties related to reciprocity, I do not claim that this justifies these elements of common morality. I am a supporter of an evolutionary approach to human behavior, and I am interested in ethics, but I am not an advocate of an “evolutionary ethic”.

The impossibility of deducing ethical conclusions from the facts of evolution does not mean that recent advances in our scientific understanding of ethics have no normative significance at all. These advances are highly significant for normative ethics, but in an indirect way. To appreciate this, we need to look at the current debate over methodology in normative ethics.

---

<sup>18</sup> Darwin to Charles Lyell, in Francis Darwin (ed.), *The Life and Letters of Charles Darwin*, vol. II (London: Murray, 1887), p. 262.

<sup>19</sup> See, for example, Edward O. Wilson, *On Human Nature* (Cambridge, MA: Harvard University Press, 1978), p. 5.

A dominant theme in normative ethics for the past century or more has been the debate between those who support a systematic normative ethical theory – utilitarianism and other forms of consequentialism have been the leading contenders – and those who ground their normative ethics on our common moral judgments or intuitions. In this debate, the chief weapons of opponents of utilitarianism have been examples intended to show that the dictates of utilitarianism clash with moral intuitions that we all share. Perhaps the most famous literary instance occurs in *The Karamazov Brothers*, where Dostoyevsky has Ivan challenge Alyosha to say whether he would consent to build a world in which people were happy and at peace, if this ideal world could be achieved only by torturing “that same little child beating her chest with her little fists.” Alyosha says that he would not consent to build such a world on those terms.<sup>20</sup> Hastings Rashdall thought he could refute hedonistic utilitarianism by arguing that it cannot explain the value of sexual purity.<sup>21</sup> H. J. McCloskey, writing at a time when lynchings in the American South were still a possibility, thought it a decisive objection to utilitarianism that the theory might direct a sheriff to frame an innocent man in order to prevent a white mob lynching half a dozen innocents in revenge for a rape.<sup>22</sup> Bernard Williams offered a similar example, of a botanist who wanders into a village in the jungle where twenty innocent people are about to be shot. He is told that nineteen of them will be spared, if only he will himself shoot the twentieth. Though Williams himself did not say that it would necessarily be wrong to shoot the twentieth, he thought that utilitarianism could not account for the difficulty of the decision.<sup>23</sup>

---

<sup>20</sup> Fyodor Dostoyevsky, *The Karamazov Brothers*, Ignat Avsey (trans.) (Oxford: Oxford University Press, 1994), Pt. 2, Bk. 5, Ch. 4.

<sup>21</sup> Hastings Rashdall, *The Theory of Good and Evil*, vol. 1 (Oxford: Clarendon Press, 1907), p. 197.

<sup>22</sup> H. J. McCloskey, “An Examination of Restricted Utilitarianism,” *Philosophical Review*, 66 (1957), pp. 466-485; reprinted in Michael D. Bayles (ed.), *Contemporary Utilitarianism* (Gloucester, MA: Peter Smith, 1978), where the example is on p. 121.

<sup>23</sup> Bernard Williams, “A Critique of Utilitarianism,” in J. J. C. Smart and Bernard Williams, *Utilitarianism; For and Against* (Cambridge: Cambridge University Press, 1973), pp. 96-100, 110-117; the relevant passage is reprinted in Peter Singer (ed.), *Ethics* (Oxford: Oxford University Press, 1994), pp. 339-345.

Initially, the use of such examples to appeal to our common moral intuitions against consequentialist theories was an ad hoc device lacking metaethical foundations. It was simply a way of saying: "If Theory U is true, then in situation X you should do Y. But we know that it would be wrong to do Y in X, therefore U cannot be true." This is an effective argument against U, as long as the judgment that it would be wrong to do Y in X is not challenged. But the argument does nothing to establish that it is wrong to do Y in X, nor what a sounder theory than U would be like. In *A Theory of Justice*, John Rawls took the crucial step towards fusing this argument with an ethical methodology when he argued that the test of a sound moral theory is that it can achieve a "reflective equilibrium" with our considered moral judgments. By "reflective equilibrium" Rawls meant that, where there is no inherently plausible theory that perfectly matches our initial moral judgments, we should modify either the theory, or the judgments, until we have an equilibrium between the two. The model here is the testing of a scientific theory. In science, we generally accept the theory that best fits the data, but sometimes, if the theory is inherently plausible, we may be prepared to accept it even if it does not fit all the data. We might assume that the outlying data are erroneous, or that there are still undiscovered factors at work in that particular situation. In the case of a normative theory of ethics, Rawls assumes, the raw data is our prior moral judgments. We try to match them with a plausible theory, but if we cannot, we reject some of the judgments, and modify the theory so that it matches others. Eventually the plausibility of the theory and of the surviving judgments reach an equilibrium, and we then have the best possible theory. On this view the acceptability of a moral theory is not determined by the internal coherence and plausibility of the theory itself, but, to a significant extent, by its agreement with those of our prior moral judgments that we are unwilling to revise or abandon. In *A Theory of Justice* Rawls uses this model to justify tinkering with his original idea of a choice arising from a hypothetical contract, until he is able to produce results that are not too much at odds with our ordinary ideas of justice.<sup>24</sup>

---

<sup>24</sup> John Rawls, *A Theory of Justice* (Cambridge, MA: Belknap Press, 1971), especially p. 48. The idea of reflective equilibrium was already present in Rawls's "Outline of a Decision Procedure for Ethics," *Philosophical Review*, 60 (1951), pp. 177-197. The analogy with a scientific theory is explicit in the earlier article.

The model of reflective equilibrium has always struck me as dubious. The analogy between the role of a normative moral theory and a scientific theory is fundamentally misconceived.<sup>25</sup> A scientific theory seeks to explain the existence of data that are about a world “out there” that we are trying to explain. Granted, the data may have been affected by errors in measurement or interpretation, but unless we can give some account of what the errors might have been, it is not up to us to choose or reject the observations. A normative ethical theory, however, is not trying to explain our common moral intuitions. It might reject all of them, and still be superior to other normative theories that better matched our moral judgments. For a normative moral theory is not an attempt to answer the question “Why do we think as we do about moral questions?” Even without an evolutionary understanding of ethics, it is obvious that the question “Why do we think as we do about moral questions?” may require a historical, rather than a philosophical, investigation. On abortion, suicide, and voluntary euthanasia, for instance, we may think as we do because we have grown up in a society that was, for the best part of 2000 years, dominated by the Christian religion. We may no longer believe in Christianity as a moral authority, but we may find it difficult to rid ourselves of moral intuitions shaped by our parents and our teachers, who were either themselves believers, or were shaped by others who were.

A normative moral theory is an attempt to answer the question “What ought we to do?” It is perfectly possible to answer this question by saying: “Ignore all our ordinary moral judgments, and do what will produce the best consequences.” Of course, one would need to give some kind of argument for this answer. My concern now is not to give this argument, or any other argument for possible alternatives to whatever theory best explains our intuitive judgments. My point is that the model of reflective equilibrium, at least as presented in *A Theory of Justice*, appears to rule out such an answer, because it assumes that our moral intuitions are some kind of data from which we can learn what we ought to do.

---

<sup>25</sup> See my “Sidgwick and Reflective Equilibrium.”

Rawls addressed the metaethical implications of his method again in *Political Liberalism*.<sup>26</sup> There he distinguished it from old-fashioned ethical intuitionism, describing it instead as “Kantian constructivism.” Whereas intuitionism seeks to defend our intuitions as offering rational insight into true ethical principles, constructivism replaces this by a search for “reasonable grounds of reaching agreement rooted in our conception of ourselves and in our relation to society”. We cannot, on this view, discover moral truth. We can only construct our moral views from concepts and ideas that we already have.

One evident objection to Rawls’s Kantian constructivism is that it makes ethics culturally relative. Different peoples, with differing conceptions of themselves and their relation to society, might construct different theories that lead them to different principles of justice. Should that be the case, it could not then be said that one set of principles is true and the other false. The most that can be claimed for the particular principles of justice that Rawls defends is that they offer reasonable grounds of agreement for people holding “our” conception of ourselves and our relation to society. But some may not see this as an objection. Cultural relativism has had many defenders in ethics, including many who misguidedly believe that it offers a defense against cultural imperialism. (This is the reverse of the truth. If ethics is culturally relative, and my culture gives great value to imposing our values on other cultures, ethical relativism allows no foothold for arguing that we are mistaken in believing that it is good to impose our values on others.) I do not, however, want to dwell on the relativist element of Kantian constructivism, because I want to make a more general objection to any method of doing ethics that judges a normative theory either entirely, or in part, by the extent to which it matches our moral intuitions.

Admittedly, it is possible to interpret the model of reflective equilibrium so that it takes into account any grounds for objecting to our intuitions, including those that I have put forward. Norman Daniels has argued persuasively for this “wide” interpretation of

---

<sup>26</sup> John Rawls, *Political Liberalism* (New York: Columbia University Press, 1993).

reflective equilibrium.<sup>27</sup> If the interpretation is truly wide enough to countenance the rejection of all our ordinary moral beliefs, then I have no objection to it. The price for avoiding the inbuilt conservatism of the narrow interpretation, however, is that reflective equilibrium ceases to be a distinctive method of doing normative ethics. Where previously there was a contrast between the method of reflective equilibrium and “foundationalist” attempts to build an ethical system outward from some indubitable starting point, now foundationalism simply becomes the limiting case of a wide reflective equilibrium.

Let us return for a moment to the trolley problem cases. As mentioned before, philosophical discussions of these cases from Judith Jarvis Thomson onwards have been preoccupied with the search for differences between the cases that justify our initial intuitive responses. If, however, Greene is right to suggest that our intuitive responses are due to differences in the emotional pull of situations that involve bringing about someone’s death in a close-up, personal way, and bringing about the same person’s death in a way that is at a distance, and less personal, why should we believe that there is anything that justifies these responses? If Greene’s initial results are confirmed by subsequent research, we may ultimately conclude that he has not only explained, but explained away the philosophical puzzle. (I say that we may ultimately reach this conclusion because of course Greene’s data alone cannot prove any normative view right or wrong. Normative argument is needed, of the kind I shall sketch below, to link those data with a particular normative view.)

This becomes clearer when we consider how well Greene’s findings fit into the broader evolutionary view of the origins of morality outlined earlier in this paper.<sup>28</sup> For most of our evolutionary history, human beings have lived in small groups, and the same is almost certainly true of our pre-human primate and social mammal ancestors. In these groups, violence could only be inflicted in an up-close and personal way – by hitting, pushing, strangling, or using a stick or stone as a club. To

---

<sup>27</sup> See Norman Daniels, *Justice and Justification: Reflective Equilibrium in Theory and Practice* (Cambridge: Cambridge University Press, 1997).

<sup>28</sup>As Greene himself has pointed out in his Ph.D. thesis, cited above.

deal with such situations, we have developed immediate, emotionally-based responses to questions involving close, personal interactions with others. The thought of pushing the stranger off the footbridge elicits these emotionally-based responses. Throwing a switch that diverts a train that will hit someone bears no resemblance to anything likely to have happened in the circumstances in which we and our ancestors lived. Hence the thought of doing it does not elicit the same emotional response as pushing someone off a bridge. So the salient feature that explains our different intuitive judgments concerning the two cases is that the footbridge case is the kind of situation that was likely to arise during the eons of time over which we were evolving; whereas the standard trolley case describes a way of bringing about someone's death that has only been possible in the past century or two, a time far too short to have any impact on our inherited patterns of emotional response. But what is the moral salience of the fact that I have killed someone in a way that was possible a million years ago, rather than in a way that became possible only two hundred years ago? I would answer: none.

Thus recent scientific advances in our understanding do have some normative significance, and at different levels. At the particular level of the analysis of moral problems like those posed by trolley cases, a better understanding of the nature of our intuitive responses suggests that there is no point in trying to find moral principles that justify the differing intuitions to which the various cases give rise. Very probably, there is no morally relevant distinction between the cases. At the more general level of method in ethics, this same understanding of how we make moral judgments casts serious doubt on the method of reflective equilibrium. There is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution as social mammals, primates, and finally, human beings. We should, with our current powers of reasoning and our rapidly changing circumstances, be able to do better than that.

A defender of the idea of reflective equilibrium might say that these arguments against giving weight to certain intuitions can themselves, on the model of “wide reflective equilibrium,” be part of the process of achieving equilibrium between a theory and our considered moral judgments. The arguments would then lead us to reject judgments that we might otherwise retain, and so end up with a different normative theory. As we have already noted, making the model of “reflective equilibrium” as all-embracing as this may make it salvageable, but only at the cost of making it close to vacuous. For with this change, the “data” that a sound moral theory is supposed to match have become so changeable that they can play, at best, a minor role in determining the final shape of the normative moral theory. Finally, for the same reasons that reflective equilibrium no longer appeals as a way of testing a moral theory, so Kantian constructivism ceases to be an attractive metaethic, whether it ends up being culturally relative or not. To the extent that “our conception of ourselves” is tied up with our intuitive ideas of right and wrong, we may question why we should be concerned to construct a moral view out of our evolved intuitions about what is the right way to act in particular situations. Moreover a Kantian constructivist who manages to avoid cultural relativism by finding universally shared intuitive ideas of right and wrong may have shown nothing more than that our common evolutionary heritage has, unsurprisingly, given us a common set of intuitive ideas of right and wrong.

What I am saying, in brief, is this. Advances in our understanding of ethics do not themselves directly imply any normative conclusions, but they undermine some conceptions of doing ethics which themselves have normative conclusions. Those conceptions of ethics tend to be too respectful of our intuitions. Our better understanding of ethics gives us grounds for being less respectful of them.

##### *5. Conclusion: A Way Forward?*

Whenever it is suggested that normative ethics should disregard our common moral intuitions, the objection is made that without intuitions, we can go nowhere. There

have been many attempts, over the centuries, to find proofs of first principles in ethics, but most philosophers consider that they have all failed. Even a radical ethical theory like utilitarianism must rest on a fundamental intuition about what is good. So we appear to be left with our intuitions, and nothing more. If we reject them all, we must become ethical skeptics or nihilists.

There are many ways in which one might try to respond to this objection, and I don't have the time here to review them all. So let me suggest just one possibility. Haidt's behavioral research and Greene's brain imaging studies suggest the possibility of distinguishing between our immediate emotionally based responses, and our more reasoned conclusions. In everyday life, as Haidt points out, our reasoning is likely to be nothing more than a rationalization for our intuitive responses – as Haidt puts it, the emotional dog is wagging the rational tail. But Greene's research suggests that in some people, reasoning can overcome an initial intuitive response. That, at least, seems the most plausible way to account for the longer reaction times in those subjects who, in the footbridge example, concluded that you would be justified in pushing the stranger in front of the trolley. These people appear to have had the same emotional responses against pushing the stranger, but further thought led them to reject that emotional response and to give a different answer. The preliminary data showing greater activity in parts of their brain associated with cognitive processes suggests the same conclusion. Moreover the answer these subjects gave is, surely, the rational answer. The death of one person is a lesser tragedy than the death of five people. That reasoning leads us to throw the switch in the standard trolley case, and it should also lead us to push the stranger in the footbridge, for there are no morally relevant differences between the two situations. (Although we may decide to withhold our praise from people who are capable of pushing someone off a footbridge in these circumstances. As Henry Sidgwick pointed out in *The Methods of Ethics*, it is important to distinguish between the utility of an action, and the utility of praising or blaming that action. We may not wish to praise those who are

capable of pushing strangers off high places, for fear that they will do it on other occasions when it does not save more lives than it costs.<sup>29)</sup>

It might be said that the response that I have called “more reasoned” is still based on an intuition, for example the intuition that five deaths are worse than one, or more fundamentally, the intuition that it is a bad thing if a person is killed. But if this is an intuition, it is different from the intuitions to which Haidt and Greene refer. It does not seem to be one that is the outcome of our evolutionary past. We have already noted Hume’s observation that “there is no such passion in human minds as the love of mankind, merely as such” and as we have seen, there is a good evolutionary reason for why this should be so. Thus the “intuition” that tells us that the death of one person is a lesser tragedy than the death of five is not like the intuitions that tell us we may throw the switch, but not push the stranger off the footbridge. It may be closer to the truth to say that it is a rational intuition, something like the three “ethical axioms” or “intuitive propositions of real clearness and certainty” to which Henry Sidgwick appeals in his defense of utilitarianism in *The Methods of Ethics*. The third of these axioms is “the good of any one individual is of no more importance, from the point of view (if I may say so) of the Universe, than the good of any other.”<sup>30</sup>

Perhaps here, after finding ourselves in broad agreement with David Hume for so much of this paper, we find the need to appeal to something in Hume’s polar opposite, Immanuel Kant. Kant thought that unless morality could be based on pure reason, it was a chimera.<sup>31</sup> Perhaps he was right. In the light of the best scientific understanding of ethics, we face a choice. We can take the view that our moral intuitions and judgments are and always will be emotionally based intuitive responses, and reason can do no more than build the best possible case for a decision already made on nonrational grounds. That approach leads to a form of moral skepticism, although one still compatible with advocating our emotionally-

---

<sup>29</sup> Henry Sidgwick, *The Methods of Ethics*, 7<sup>th</sup> ed. (London: Macmillan, 1907), pp. 428-429.

<sup>30</sup> Sidgwick, *The Methods of Ethics*, p. 382.

<sup>31</sup> Immanuel Kant, *Groundwork of the Metaphysics of Morals*, trans. Mary Gregor, Cambridge University Press, Cambridge, 1997, Sec. II.

based moral values and encouraging clear thinking about them.<sup>32</sup> Alternatively, we might attempt the ambitious task of separating those moral judgments that we owe to our evolutionary and cultural history, from those that have a rational basis. This is a large and difficult task. Even to specify in what sense a moral judgment can have a rational basis is not easy. Nevertheless, it seems to me worth attempting, for it is the only way to avoid moral skepticism.

---

<sup>32</sup> Greene takes this position in *The Terrible, Horrible, No Good, Very Bad Truth about Morality, and What to Do About It*. He describes his view as moral skepticism, but distinguishes it from moral nihilism, in which there is no place for moral values at all. I am grateful to Greene, not only for his illuminating research, but for his valuable comments on this paper. Other helpful comments have come from people too numerous to mention individually, so I offer collective thanks to all those who spoke up when I presented this paper at the James Rachels Memorial Conference at the University of Alabama, Birmingham; at the Princeton University Center for Human Values Fellows' Seminar; and at Philosophy Departments at the following universities: University of Melbourne, University of Vermont, Rutgers University, and the University of Lodz. Despite all this advice, I am aware that there is much more work needed: this paper is no more than a sketch of an argument that I hope to develop more adequately in future.