The Status of Moral Emotions in Consequentialist Moral Reasoning

By

Robert H. Frank

Cornell University

Presented at "Free Enterprise: Values in Action"

Gruter Institute/John Templeton Foundation/

UCLA-Sloan Research Program

4th Working Conference

at

Harvard Business School

June 20-21, 2006

The Status of Moral Emotions in Consequentialist Moral Reasoning

by

Robert H. Frank[1]

The philosopher Bernard Williams describes an example in which a botanist wanders into a village in the jungle where ten innocent people are about to be shot. He is told that nine of them will be spared if he himself will shoot the tenth. What should the botanist do? Although most people would prefer to see only one innocent person die rather than ten, Williams argues that killing the innocent villager is not by any means the obviously right thing for Jim to do.[2] And most people seem to agree.

The force of the example is its appeal to a widely shared moral intuition. Yet some philosophers counter that it is the presumed validity of moral intuitions that such examples call into question (Singer, 2002). These *consequentialists* insist that whether an action is morally right depends only on its consequences. The right choice, they argue, is always the one that leads to the best overall consequences.

I will argue that consequentialists make a persuasive case that moral intuitions are best ignored in at least some specific cases. But many consequentialists appear to take the stronger position that moral intuitions should play no role in moral choice. I will argue against that position on the grounds that should appeal to their way of thinking. As I will attempt to explain, ignoring moral intuitions would lead to undesirable consequences. My broader aim is to expand the consequentialist framework to take explicit account of moral sentiments.

---

[1] Henrietta Johnson Louis Professor of Management and Professor of Economics, Johnson Graduate School of Management, Cornell University, Ithaca, NY 14853.

[2] Williams, 1973.

Consequentialist vs. Deontological Moral Theories

Consequentialism differs from traditional, or *deontological*, moral theories, which hold that the right choice must be identified on the basis of underlying moral principles. These principles may spring from religious tradition (for example, the Ten Commandments) but need not (for example, Kant's categorical imperative). Whatever their source, the moral force of the principles invoked by deontologists increases with the extent to which these principles accord with strongly held moral intuitions.

For many, perhaps even the overwhelming majority, of cases, consequentialist and deontological moral theories yield the same prescriptions. Both camps, for example, hold that it was wrong for Enron executives to have lied about their company's earnings and wrong for David Berkowitz to have murdered six innocent people. Even in cases in which there might appear to be ample room for disagreement about what constitutes moral behavior, a majority of practitioners from both camps often take the same side.

Consider, for example, a variant of the familiar trolley-car problem discussed by philosophers. You are standing by a railroad track when you see an out-of-control trolley car about to strike a group of five people standing on the tracks ahead. You can throw a nearby switch, diverting the trolley onto a side track, which would result in the death of one person standing there. Failure to throw the switch will result in all five persons being killed on the main track.

Consequentialists are virtually unanimous in concluding that the morally correct choice is for you to throw the switch. Some deontologists equivocate, arguing that the active step of throwing the switch would make you guilty of killing the person on the side

track, whereas you would not be guilty of killing the five on the main track if you failed to intervene. Yet even most deontologists conclude that the distinction between act and omission is not morally relevant in this example, and that your best available choice is to throw the switch.

But even though the two moral frameworks exhibit broad agreement with respect to the ethical choices we confront in practice, many deontologists remain deeply hostile to the consequentialist framework.

The Status of Moral Intuitions

Critics often attack consequentialist moral theories by constructing examples in which the choice that consequentialism seems to prescribe violates strongly held moral intuitions. In another version of the trolley-car problem, for example, the trolley is again about to kill five people, but this time you are not standing near the tracks but on a footbridge above them. There is no switch you can throw to divert the train. But there is a large stranger standing next to you, and if you push him off the bridge onto the tracks below, his body will derail the trolley, in the process killing him but sparing the lives of the five strangers. (It won't work for you to jump down onto the tracks yourself, because you are too small to derail the trolley.)

Consequentialism seems to prescribe pushing the large stranger from the bridge, since this would result in a net savings of four lives. Yet when people are asked what they think should be done in this situation, most feel strongly that it would be wrong to push the stranger to his death. Those who share this intuition are naturally sympathetic to the deontologists' claim that the example somehow demonstrates a fundamental flaw in

the consequentialist position. This version of the trolley problem thus elicits essentially the same moral judgment as Bernard Williams's example involving the botanist.

Many consequentialists, Princeton philosopher Peter Singer among them, question the validity of the moral intuitions evoked by such examples (Singer, 2002). To illustrate, Singer asks us to imagine another variant of the trolley problem, one that is identical to the first except for one detail. You can throw a switch that will divert the train not onto a side track, but onto a loop that circles back onto the main track. Standing on the loop is a large stranger whose body would bring the trolley to a halt if it were diverted onto the loop. Singer notes that this time most people say that the right choice is to divert the trolley, just as in the original example in which the switch diverted the trolley onto a side track rather than a loop. In both cases, throwing the switch caused the death of one stranger, in the process sparing the lives of the five others on the main track.[3]

Singer's Princeton colleague Joshua Greene, a cognitive neuroscientist, has suggested that people's intuitions differ in these two examples not because the morally correct action differs, but rather because the action that results in the large stranger's death is so much more vivid and personal in the footbridge case than in the looped-track case:

> 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

[3] The looped-track example suggests that it was not the Kantian prohibition against using people merely as means that explains the earlier reluctance to push the stranger from the footbridge, since choosing to throw the switch in the looped-track example also entails using the stranger as merely a means to save the other five. In the original example, diverting the trolley onto the side track would have saved the others even if the stranger had not been on the side track.

> them to say that the action in the original case is permissible. (Greene, 2002, p. 178.)

To test this explanation, Green used functional magnetic resonance imaging to examine activity patterns in the brains of subjects confronted with the two decisions. His prediction was that activity levels in brain regions associated with emotion would be higher when subjects considered pushing the stranger from a footbridge than when they considered diverting the trolley onto the looped track. He also reasoned that the minority of subjects who felt the right action was to push the stranger from the footbridge would reach that judgment only after overcoming their initial emotional reactions to the contrary. Thus he also predicted that the decisions taken by these subjects would take longer than those reached by the majority who thought it wrong to push the stranger to his death, and longer as well than it took for them to decide what to do in the looped-track example. Each of these predictions was confirmed.

Is it morally relevant that thinking about causing someone's death by pushing him from a footbridge elicits stronger emotions than thinking about causing his death by throwing a switch? Peter Singer argues that it is not—that the difference is a simple, non-normative consequence of our evolutionary past. Under the primitive, small-group conditions under which humans evolved, he argues, the act of harming others always entailed vivid personal contact at close quarters. One could not cause another's death by simply throwing a switch. So if it was adaptive to be emotionally reluctant to inflict harm on others—surely a plausible presumption—the relevant emotions ought to be much more likely to be triggered by vivid personal assaults than by abstract actions like throwing a switch.

A historical case in point helps highlight the distinction. Shortly after British intelligence officers had broken Nazi encryption schemes in World War II, Winston Churchill had an opportunity to spare the lives of British residents of Coventry by warning them of a pending bombing attack. To do so, however, would have revealed to the Nazis that their codes had been broken. In the belief that preserving the secret would save considerably more British lives in the long run, Churchill gave Coventry no warning, resulting in large numbers of preventable deaths.

It is difficult to imagine a more wrenching decision, and we celebrate Churchill's moral courage in making it. But it is also easy to imagine that Churchill would have chosen differently had it been necessary for him personally to kill the Coventry's residents at close quarters, rather than merely to allow their deaths by failing to warn them. Singer's claim is that while this difference is a predictable consequence of the way in which natural selection forged our emotions, it has no moral significance.

In sum, the fact that consequentialist moral theories sometimes prescribe actions that conflict with moral intuitions cannot, by itself, be taken as evidence against these theories. Moral intuitions are contingent reactions shaped by the details of our evolutionary history. Often they will not be relevant for the moral choices we confront today.

Moral Sentiments as Commitment Devices

The fact that it might sometimes be best to ignore moral emotions does not imply that it is always, or even usually, best to ignore them. If we are to think clearly about the role of moral emotions in moral choice, we must consider the problems that these

emotions were molded by natural selection to solve.

Most interesting moral questions concern actions the individual would prefer to take except for the possibility of causing undue harm to others. Unbridled pursuit of self-interest often results in worse outcomes for everyone. In such situations, an effective moral system curbs self-interest for the common good.

But as Handy, O'Hara, and Solomon have emphasized elsewhere in this volume, moral systems must not only identify which action is right, they must also provide motives for taking that action. The difficulty of serving both goals at once is immediately apparent. Humans evolved in harsh environments in which the consequences of failure to pursue self-interest were often severe. Famines and other threats to survival were common. Polygyny was also the norm in early human societies, which for men meant that failure to achieve high rank ensured failure to marry. Under the circumstances, it was an obvious challenge to motivate individuals to forgo self-interest for the common good.

Yet instances in which people forgo self-interest are actually quite common, even though the human nervous system is much the same today as it was tens of thousands of years ago. For example, although self-interest dictates leaving no tip in restaurants you don't expect to visit again, most people tip at about the same rate at such restaurants as in local restaurants.[4] Similarly, when sociologists perform the experiment of dropping wallets containing small amounts of cash on sidewalks in New York, about half are returned by mail with the cash intact.[5]

---

[4] Bodvarsson and Gibson, 1994. Tipping in local restaurants can be rationalized as a self-interested activity: if you don't tip well, you might not get good service the next time. People resist the temptation to stiff the waiter because the shadow of the future is staring at them

[5] Hornstein, 1976.

The Falklands War is another good example. The British could have bought the Falklanders out—giving each family, say, a castle in Scotland and a generous pension for life—for far less than the cost of sending their forces to confront the Argentines. Instead they incurred considerable cost in treasure and lives. Yet few in the UK opposed the decision to fight for the desolate South Atlantic islands. One could say that Margaret Thatcher gained political by responding as she did, but this begs the question of why voters preferred retaliation to inaction. When pressed, most people speak in terms of the nation's pride being at stake.

People rescue others in distress even at great peril to themselves; they donate bone marrow to strangers. Such actions are in tension with the self-interest model favored by economists. They seem to be motivated by moral sentiments. But where do these sentiments come from?

Adam Smith said that moral sentiments were endowed in us by the creator for the good of society. It is true that society works better if moral sentiments motivate people to exercise restraint. But as Darwin emphasized, selection pressures are generally far weaker at the society level than at the level of the individual organism. Moral sentiments often motivate people to incur costs that they could avoid. On what basis might these sentiments have been favored by natural selection?

In my 1988 book, *Passions Within Reason*, I proposed a mechanism based on Tom Schelling's work on the difficulties people face when confronted with *commitment problems*.[6] He illustrates the basic idea with an example of a kidnapper who seizes a victim and then gets cold feet. The kidnapper wants to set the victim free but knows that, once freed, the victim will reveal the kidnapper's identity to the police. So the kidnapper

[6] Schelling, 1960.

reluctantly decides he must kill the victim. In desperation, the victim promises not to go to the police. The problem is that both know that once he is out the door, his motive for keeping that promise will vanish.

Schelling suggests the following ingenious solution: If there is some evidence of a crime that the victim has committed, he can share that evidence with the kidnapper, which will create a bond ensuring his silence. The evidence of the victim's crime is a commitment device that makes an otherwise empty promise credible.

Schelling's basic insight can be extended to show why a trustworthy person might be able to prosper even in highly competitive market settings. Suppose you have a business that is doing so well that you know it also would thrive in a similar town 300 miles away. The problem is that because you can't monitor the manager who would run this business, he would be free to cheat you. Suppose a managerial candidate promises to manage honestly. You must then decide whether to open the branch outlet. If you do and your employee manages honestly, you each come out very well—say, $1,000 each better than the status quo. But if the manager cheats, you will lose $500 on the deal, and he will gain $1,500 relative to the status quo. The relevant payoffs for the various options are thus as summarized in Figure 1. (Note that these payoffs define a trust game like the ones discussed elsewhere in this volume by Schwab and Ostrom and by Bergstrom, Kerr, and Lachman.)

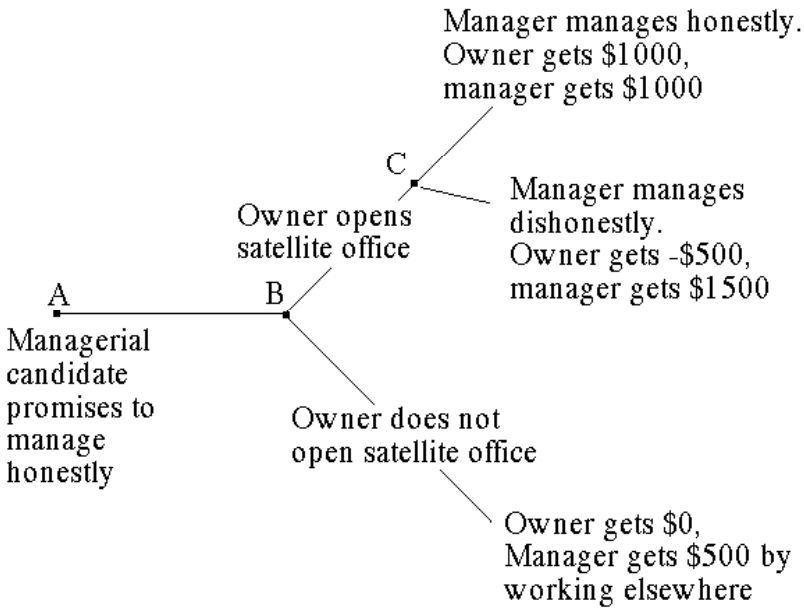

Figure 1. The Branch-Outlet Problem

If you open the outlet, the manager finds himself on the top branch of the decision tree, where he faces a choice between cheating and not. If he cheats his payoff is $1,500; if not, his payoff is only $1,000. Standard rational choice models assume that managers in these situations will be self-interested in the narrow sense. If that is your belief, you predict that the manager will cheat, which means your payoff will be -$500. And since that is worse than the payoff of zero you would get if you didn't open the branch outlet, your best bet is not to open it. The pity is that this means a loss to both you and the manager relative to what could have been achieved had you opened the branch outlet and the manager run it honestly.

Now suppose you can identify a managerial candidate who would be willing to pay $10,000 to avoid the guilt he would feel if he cheated you. Needless to say, using a financial penalty as a proxy for guilt feelings would be inappropriate in normative discourse. We would not say, for example, that it's OK to cheat as long as you gain

enough to compensate for the resulting feelings of guilt. But the formulation does nonetheless capture an important element of behavior. Incentives matter, and people are less likely to cheat when the penalties are higher.

In any event, it is clear how this simple change transforms the outcome of the game. If the manager cheats, his payoff is not $1,500 but –$8,500 (after the $10,000 psychological burden of cheating is deducted). So if you open the branch outlet, the manager will choose to manage honestly, and both he and you come out ahead. If you could identify a trustworthy manager in this situation, he or she would not be at a disadvantage. On the contrary, both you and that manager would clearly profit.

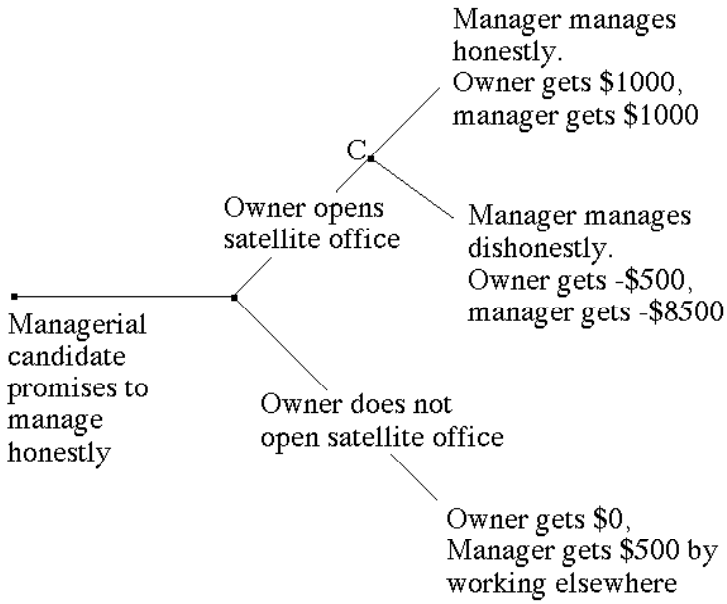

Figure 2. The Branch-Outlet Problem with an Honest Manager

Note, however, that the managerial candidate won't be hired unless his taste for honesty is observable. Thus an honest candidate who is not believed to be honest fares

worse than a dishonest candidate is believed to be honest. The first doesn't even get hired. The second not only gets the job but also the fruits of cheating the owner.

Imagine a mutation that caused trustworthy people to be born with an identifying mark, such as a 'c' on their foreheads (for 'cooperator'). Then the problem would be solved. In such a world, the trustworthy types would drive the untrustworthy types to extinction. If the two types were costlessly distinguishable, the only equilibrium would be one with pure trustworthy types in the population.[7]

In general, though, it is costly to figure out who is trustworthy. So people would not vigilant in their choice of trading partners in an environment in which everyone was trustworthy. Being vigilant would not pay, just as buying an expensive security system for your apartment would not pay if you lived in a neighbourhood in which there had never been a burglary. Thus a population consisting exclusively of trustworthy types could not be an evolutionarily stable equilibrium. Given reduced levels vigilance, untrustworthy types could easily invade such a population. So if character traits are costly to observe, the only sustainable equilibrium is one in which there is a mixed population consisting of both honest and dishonest individuals.

How might a signal of trustworthiness have emerged in the first place? Even if the first trustworthy person bore some observable marker, no one else would have had any idea what it meant. Nico Tinbergen argued that a signal of any trait must originate completely by happenstance.[8] That is, if a trait is accompanied by an observable marker, the link between the trait and the marker had to have originated by chance. For example, the dung beetle escapes predators by resembling the dung on which it feeds. How did it

[7] Frank, 1988, chapter 3.
[8] Tinbergen, 1952.

get to look like this? Unless it just happened to look enough like a fragment of dung to have fooled the most near-sighted predator, the first step toward a more dunglike appearance couldn't have been favoured by natural selection. As Stephen Jay Gould asked, "Can there be any advantage in looking 5 percent like a turd?"[9] The problem is that no predator would be fooled. So how is the threshold level of resemblance reached? It must begin with a purely accidental link between appearance and surroundings. But once such a link exists, then selection can begin to shape appearance systematically.

Similarly, we may ask, "How could a moral sentiment have emerged if no one initially knew the significance of its accompanying marker?" One hypothesis is suggested by the logic of the iterated prisoner's dilemma. There is no difficulty explaining why a self-interested person would cooperate in an iterated prisoner's dilemma.[10] If you are a tit-for-tat player, for example, and happen to pair with another such player on the first round, you and that other player will enjoy an endless string of mutual cooperation.[11] For this reason, even Attila the Hun, lacking any moral sentiments, would want to cooperate on the first move of an iterated prisoner's dilemma. The problem is that if you cooperate on the first move, you forgo some gain in the present moment, since defection on any iteration always yields a higher payoff than cooperation. It is well known that both humans and other animals tend to favour small immediate rewards over even much larger long-term rewards.[12] So, even though you expect to more than recoup the immediate sacrifice associated with cooperation, you may discount those future gains excessively. Successful cooperation, in short, requires self-control.

---

[9] Gould, 1977, p. 104.
[10] Frank, 1988, chapter 4.
[11] Rapoport and Chammah, 1965.
[12] Ainslie, 1992.

If you were endowed with a moral sentiment that made you feel bad when you cheated your partner, even if no one could see that you had that sentiment, this would make you better able to resist the temptation to cheat in the first round. And that, in turn, would enable you generate a reputation for being a cooperative person, which would be clearly to your advantage.

Moral emotions may thus have originated as impulse-control devices. This interpretation accords with observations elsewhere in this volume by Bergstrom, Kerr, and Lachman, who argue, in effect, that a person's willingness to "waste" time in social relationships may serve as a commitment device. In their account, willingness to waste time would be a relatively costly step for defectors, who would be forced to seek other relationships anew if discovered cheating. My argument suggests a complementary interpretation: An inclination to spend seemingly unproductive time in social relationships may be also be productive because it signals capacities to experience sympathy or empathy, which also make cheating more costly.

In any event, the activation of these emotions, like other forms of brain activation, may be accompanied by involuntary external symptoms that are observable. If so, the observable symptoms over time could have become associated in others' minds with the presence of these moral sentiments. And once that association was recognized, the moral emotions would be able to play a second role—namely, that of helping people solve one-shot prisoner's dilemmas and other commitment problems. The symptoms themselves can then be further refined by natural selection because of their capacity to help identify reliable partners in one-shot dilemmas.

How do you communicate something to another individual who has reason to be skeptical of what you are saying? Suppose, for example, that a toad meets a rival and both want the same mate. Among animals generally, the smaller of two rivals defers to the larger, thereby avoiding a costly fight that he would have been likely to lose anyway. Rival toads, however, often encounter one another at night, making visual assessment difficult. What they do is croak at one another, and the toad with the higher-pitched croak defers. The idea is that the lower your croak, the bigger you are on average. So it is prudent to defer to the lower croaker. This example illustrates the costly-to-fake principle: "I'll believe you not because you *say* you are a big toad, but rather because you are using a signal that is difficult to present unless you really *are* a big toad."[13]

It is the same when dogs face off: they seem to follow an algorithm of deferring to the larger dog. Consider the drawings in Figure 3, taken from Charles Darwin's 1872 book, *The Expression of Emotion in Man and Animals*. The left panel portrays a dog that is confronting a rival. Darwin argued that we reliably infer what is going on emotionally in this dog's brain by observing the numerous elements of its posture are so serviceable in the combat mode: The dog's hackles are raised, its fangs are bared, its ears are pricked, its eyes wide open and alert, its body poised to spring forward. Darwin reasoned that any dog that had to go through a conscious checklist to manifest these postural elements one by one would be too slow on the draw to compete effectively against a rival in whom the entire process was activated autonomously by the relevant emotional arousal. That autonomous link, he concluded, provides a window into the dog's brain.

[13] Frank, 1988, chapter 6.

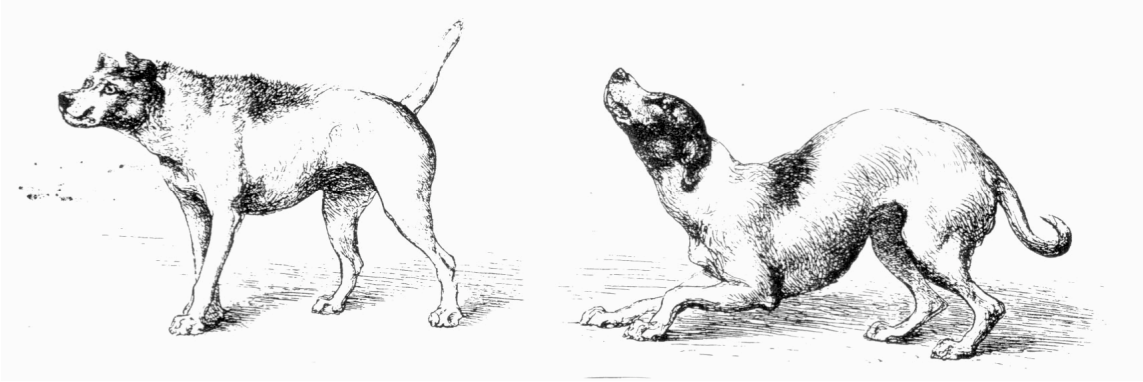

Figure 3. Observable Correlates of Emotional States

Darwin argued that there are similar emotional symptoms in humans.[14] Certain expressions, for example, spring unbidden to the human face in the presence of triggering emotions, yet are extremely difficult to present when those emotions are absent. People raised in different cultural traditions around the world can readily identify the schematic expression portrayed in Figure 4 as one corresponding to emotions such as sadness or concern. As Paul Ekman and his colleagues have shown, most people are unable to reproduce this expression on command.[15] Various other emotions also have their characteristic signatures.

[14] Darwin, 1872.
[15] Ekman, 1985.

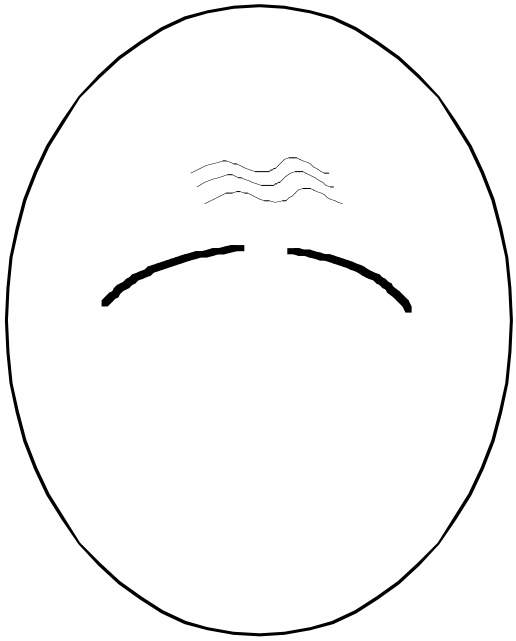

Figure 4. The Characteristic Expression of Sadness or Concern

In the argument I am attempting to advance, emotions like sympathy and empathy play a central role. Other authors in this volume have also stressed a similar role these emotions. Brosnan and de Waal, for example, report that precursors of these emotions are clearly visible in some primates and appear to motivate sharing. Solomon discusses the illustrious history of these emotions in moral discourse, beginning with David Hume and Adam Smith. And Zak notes the relationship between the neuroactive hormone oxytocin levels and the experience of empathy and sympathy. Those of us who focus on these emotions stress their role in motivating individuals to forgo gain in deference to the interests of others. My particular concern is with the question of how a person might identify the presence of these emotions in others.

How, in other words, can you predict whether someone's sympathy for you will prevent him from cheating you, even when he has an opportunity to do so with no

possibility of being punished? Clearly would not suffice for him merely to say that he was sympathetic to your interests. You need a more reliable signal. In the branch-outlet problem discussed earlier, for example, most people would feel more comfortable hiring an old friend than a perfect stranger.

I have been talking thus far as if there were good people and bad people, with some mixture of these two pure types comprising the entire population. Life is of course more complicated: we have all done something decent, and we have all cheated at one point or another. The question is, under what conditions do we cheat? Evidence suggests that we are far less likely to cheat others with whom we enjoy strong sympathetic bonds.

Such bonds appear to form as a result of a complex dance that plays out among people over time.[16] When you have a commitment problem to solve, you pick somebody who you think cares about you. People are critical of George W. Bush for giving jobs to cronies. But all leaders do that, and for good reason. Bush's particular problem has been that many of his cronies were not competent. The idea that you would pick someone well known to you is intelligible: this is a sensible thing to do. In the end, the question is whether we can identify who will cheat and who won't.

Tom Gilovich, Dennis Regan and I have done some experiments on this.[17] Our subjects had conversations in groups of three for 30 minutes, at the end of which time they played prisoner's dilemma games with each of their conversation partners. Subjects were sent to separate rooms to fill out forms on which they indicated, for each partner, whether they were going to cooperate or defect. They also recorded their predictions of what each partner would do when playing with them. Each subject's payoff was the

[16] For a rich description, see Sally, 2000.
[17] Frank, Gilovich, and Regan, 1993.

calculated as the sum of the payoffs from the relevant cells of the two games, plus a random term, so no one knew after the fact who had done what.

Almost 74 percent of the people cooperated in these pure one-shot prisoner's dilemmas. This finding is completely unpredicted by the standard self-interest model. But other empirical studies have also found high cooperation rates in dilemmas when subjects were allowed to communicate.[18] Our particular concern was with whether subjects could predict how each of their specific partners would play. When someone predicted that a partner would cooperate there was an 81 percent likelihood of cooperation (as opposed to the 74 percent base rate). On the defection side, the base rate was just over 26 percent, but partners who were predicted to defect had a defection rate of almost 57 percent. This seems an astonishingly good prediction on the basis of just 30 minutes of informal conversation.

The following thought experiment also speaks to the question of whether people can make accurate predictions about who will cheat them. Imagine that you have just returned from a crowded concert to discover that you have lost $1,000 in cash. The money, which had been in an envelope with your name and address on it, apparently fell from your coat pocket while you were at the concert. Do you know anyone not related to you by blood or marriage who you feel certain would return your money? Most people say they do. What makes them feel so confident?

Note that it is extremely unlikely that they have experienced this situation before. But even if they had, if the person named had found their money and kept it, they would not have known that. Under the circumstances, returning the money is a strict contradiction of the narrow self-interest model favored by economists.

[18] See Sally, 1995.

Many people find it natural to say that the act of returning the money in a situation like this must be motivated by some sort of moral emotion. Thus, people might predict that a friend would return their money because the friend would feel bad about the prospect of keeping it.

How did you pick the person who you thought would return your money if she found it? Typically it is someone with whom you have developed sympathetic bonds over an extended period. The feeling is that you know enough about this person to say that if she found your money, she wouldn't feel right about keeping it.

To say that trustworthiness could be an evolutionarily stable strategy is not to say that everyone is primed to cooperate all the time. Opportunism of the sort predicted by self-interest models is in abundant supply. Yet the prospects for sustaining cooperation are not as bleak as many economists seem to think. Many people are willing to set aside self-interest to promote the common good. Even if moral emotions are unobservable by others, they can still help you to be patient in repeated prisoner's dilemmas. But if others recognize you to be a decent person, there are all sorts of ways in which you are valuable. If you are in business, your boss is likely to have a firm opinion about whether you'd be the sort of person to return the lost $1,000 if you found it. You'd like him to think that you'd return it. The best way to get him to think that, it appears, is to actually be the kind of person who would return it.

Do Our Models of Human Nature Matter?

As Gintis and Khaurana note elsewhere in this volume, neoclassical economic models typically assume that people are self-interested in the narrow sense. Yet abundant

evidence suggests that motives other than self-interest are also important. An obvious consequence of inaccurate behavioral assumptions is that they often lead to inaccurate predictions. But in their papers in this volume, Casebeer and Stout note another worrisome possibility—namely, that inaccurate portrayals of human nature may prove self-reinforcing.

The self-interest model of rational choice predicts that people will defect in one-shot prisoner's dilemmas. Does working with that model over the course of many years, as professional economists do, alter their expectations about what others will do in social dilemmas? And if so, does this alter how economists themselves behave when confronted with such dilemmas? Tom Gilovich and Dennis Regan and I found that economics training—both its duration and content—affects the likelihood that undergraduate students will defect in prisoner's dilemma games.[19] In one version of our experiments, economics majors were almost twice as likely to defect as non-economics majors. This difference could stem in part from the fact that people who elect to study economics were different from others in the first place. But we also found at least weak evidence for the existence of a training effect. The differences in cooperation rates between economics majors and non-majors increased with the length of time that students had been enrolled in the major. We also found that, relative to a control group of freshmen astronomy students, students in an introductory microeconomics course were more likely in December than in September to expect opportunistic behavior on the part of others.

[19] Frank, Gilovich, and Regan, 1993a; Marwell and Ames (1981) and Carter and Irons (1991) report similar findings.

My point is not that my fellow economists are wrong to stress the importance of self-interest. But those who insist that it is the only important human motive are missing something important. Even more troubling, the narrow self-interest model, which encourages us to expect the worst in others, may bring out the worst in us as well.

Difficulties Confronting Emotion-Free Consequentialism

Consequentialist moral systems that ignore moral emotions face multiple challenges. It is one thing to say that we would all enjoy greater prosperity if we refrained from cheating one another. But it is quite another to persuade individuals not to cheat when cheating cannot be detected and punished.

Even for persons strongly motivated to do the right thing, consequentialist moral systems can sometimes make impossible demands on individuals. Imagine, for example, that five strangers are about to be killed by a runaway trolley, which at the flip of a switch you could divert onto a side track where it would kill four of your closest friends. Many consequentialists would argue that it is your moral duty to flip the switch, since it is better that only four die instead of five. But a person capable of heeding such advice would have been unlikely to have had any close friends in the first place. Indeed, it is easy to imagine that most people would become more reluctant to form close friendships if they believed it their duty to ignore the emotional bonds that such friendships inevitably entail.

The capacity to form deep bonds of sympathy and affection is important for solving a variety of commitment problems. It is not a capacity easily abandoned. And even if we could abandon it, the emotional and material costs would be substantial.

Do Moral Emotions Define Right Conduct?

Since our current environment differs in many important ways from the environments in which our ancestors evolved, we should not be surprised that our intuitions sometimes mislead us about today's moral questions. Thus we have not just Singer's example of an inhibition that is too strong (our reluctance to push the fat man from the bridge to save the five strangers), but also many others in which our inhibitions are too weak (such as those against stealing from corporate employers, filing overstated insurance claims, or understating our incomes for tax purposes). In the latter examples, the weakness of inhibition is plausibly explained by the fact that in the environments in which we evolved, cheating always victimized specific persons rather than faceless institutions.

But our moral intuitions don't always mislead us. In the lost-envelope thought experiment, for example, my misgivings about keeping my friend's cash would push me to do what an independent consequentialist moral analysis says I ought to do under the circumstances—namely, return the cash. Indeed, our intuitions appear to provide sound moral guidance more often than not. For this reason, taking them at face value seems like a reasonable default option, provided we remain open to the possibility that they may be misleading in specific cases.

That said, I must emphasize that my argument about the moral emotions in

Passions Within Reason was intended to serve one purpose only—to explain how people who evolved under the pressures of natural selection might nonetheless have inherited motivations to do the right thing under some circumstances in which such conduct entailed avoidable costs. I never claimed that our intuitions define right conduct.

As noted, the only equilibria that are sustainable in the evolutionary games I discuss entail populations containing at least some individuals who lack the usual moral inhibitions. We must incur costs to engage in character assessment, and it would make no sense to incur these costs if everybody were inclined to do the right thing all the time. But if people were never vigilant when choosing their trading partners, then mutant cheaters could invade an honest population at will. And since any population must therefore contain at least some cheaters in equilibrium, a given individual's moral intuitions simply cannot be used to define what constitutes right conduct. That assessment requires an independent analysis based on some sort of moral theory.

Are our moral intuitions relevant to the choice of which moral theory to employ? In some cases, absolutely yes. But in at least some others, I agree with Singer that we must be prepared to embrace a moral theory even though it might conflict with a specific moral intuition we hold dear.

If I read him correctly, however, he goes too far in claiming that moral intuitions should play no role at all in moral judgment—either in choosing among moral theories or in performing moral analysis within the framework of any given theory. Since our moral intuitions are in harmony with our moral theories most of the time, this claim seems strange on its face. (It is of course consistent with his contrarian nature!) Singer's point, though, is that the apparent harmony is less informative than it seems, because the

authors of moral theories consciously strive to make them consistent with our intuitions. Fair enough, but that clearly does not imply that moral intuitions are generally irrelevant.

On the contrary, since they appear to provide useful guidance more often than not, we should be prepared to offer a coherent account for why a given intuition is misleading before proposing to dismiss it. That strategy works just fine in specific cases. For instance, it seems plausible to explain our relative lack of inhibition against using weapons that kill at great distances (and, by extension, our lack of inhibition about killing a stranger by flipping a trolley switch) by saying that killing in such remote ways simply wasn't possible in the ancestral environment.

But to say we should disregard moral emotions generally, one would have to offer a similar argument against each of them. And this, I believe, Singer cannot do. It is for this reason that I find him unpersuasive when he insists that it is necessarily better to save two strangers than a single friend. His claim violates a strongly held intuition, but this time it is one that cannot easily be shown to be misleading.

Concluding Remarks

In brief, I have argued for a middle ground between and Singer's position and that of John Rawls, who argued that progress in moral theory results from efforts to reconcile our moral theories with our moral intuitions.[20] Insofar as I believe that a moral theory is likely to be judged unacceptable if it systematically violates our moral intuitions, I am more or less on Rawls's side. But with Singer, I am prepared to embrace a moral theory that violates a specific moral intuition if a plausible account can be given for why that intuition is misleading.

[20] Rawls, 1971.

References

Ainslie, George (1992). *Picoeconomics*, New York: Cambridge University Press.

Bodvarsson, O. B. and W. A. Gibson (1994). Gratuities and customer appraisal of service: evidence from Minnesota restaurants," *Journal of Socioeconomics*, 23, 1994: 287-302.

Carter, John and Michael Irons (1991). "Are Economists Different, and If So, Why?" *Journal of Economic Perspective,* 5, Spring.

Darwin, Charles (1965; 1872). *The Expression of Emotions in Man and Animals*. Chicago: University of Chicago Press.

Ekman, Paul (1985). *Telling Lies*. New York: W. W. Norton.

Frank, Robert H. (1988). *Passions Within Reason*. New York: W. W. Norton.

Frank, Robert H., Gilovich, Thomas, and Regan, Dennis (1993a). Does studying economics inhibit cooperation? *Journal of Economic Perspectives, 7,* Spring, 159-171.

_____ (1993b). The evolution of one-shot cooperation. *Ethology and Sociobiology*, *14,* July, 247-256.

Gould, Stephen Jay (1977). *Ever Since Darwin*. New York: W. W. Norton.

Greene, Joshua. *The Terrible, Horrible, No Good, Very Bad Truth about Morality, and What to Do About It,* Department of Philosophy, Princeton University, 2002.

Greene J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment" *Science*, 293 (5537), 2105-8.

Hornstein, Harvey (1976). *Cruelty and Kindness*, Englewood Cliffs, NJ: Prentice Hall.

Hume, David (1978; 1740). *A Treatise of Human Nature*. Oxford: Oxford University Press.

Marwell, Gerald and Ruth Ames (1981). "Economists Free Ride, Does Anyone Else?" *Journal of Public Economics* 15: 295-310.

Rapoport, Anatol and A. Chammah (1965). *Prisoner's Dilemma.* Ann Arbor: University of Michigan Press.

Rawls, John (1971). *A Theory of Justice*, Cambridge, Massachusetts: Harvard Belknap, 1971.

Sally, David (1995). Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1972. *Rationality and Society, 7, 58-92*.

_____ (2000). A general theory of sympathy, mind-reading, and social interaction, with an application to the prisoners' dilemma. *Social Science Information*, 39(4), 567-634.

Schelling, Thomas C. (1960). *The Strategy of Conflict*, New York: Oxford University Press.

Singer, Peter. "The Normative Significance of Our Growing Understanding of Ethics," paper presented at the Ontology Conference, San Sebastian, Spain, October 3, 2002.

Smith, Adam (1966; 1759). *The Theory of Moral Sentiments*. New York: Kelley.

Tinbergen, Niko (1952). Derived activities: their causation, biological significance, and emancipation during evolution." *Quarterly Review of Biology* 27: 1-32.

Williams, Bernard (1973). "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, eds., Utilitarianism: For and Against, Cambridge, Cambridge University Press.