Philosophical Explorations, Vol. 9, No. 1, March 2006

Routledge
Taylor & Francis Group

THE EMOTIONAL BASIS OF MORAL JUDGMENTS

Jesse Prinz

Recent work in cognitive science provides overwhelming evidence for a link between emotion and moral judgment. I review findings from psychology, cognitive neuroscience, and research on psychopathology and conclude that emotions are not merely correlated with moral judgments but they are also, in some sense, both necessary and sufficient. I then use these findings along with some anthropological observations to support several philosophical theories: first, I argue that sentimentalism is true: to judge that something is wrong is to have a sentiment of disapprobation towards it. Second, I argue that moral facts are response-dependent: the bad just is that which cases disapprobation in a community of moralizers. Third, I argue that a form of motivational internalism is true: ordinary moral judgments are intrinsically motivating, and all non-motivating moral judgments are parasitic on these.

Introduction

In the early 1970s, social psychologist Stanley Milgram instructed his graduate students to approach strangers on a New York City subway and request their seats. Almost all of Milgram's students refused to try this, and the one student who was willing to go came back quickly and reported that he had to abort the experiment before collecting enough data. The student had not been physically threatened in any way, and indeed the majority of people he asked willingly gave up their seats. Milgram couldn't understand why his student came back prematurely, and he decided to descend into the subway and perform the experiment himself. This is how he recalls his experience:

> The words seemed lodged in my trachea and would simply not emerge. Retreating, I berated myself: 'What kind of craven coward are you?'
>
> Finally after several unsuccessful tries, I went up to a passenger and choked out the request, 'Excuse me sir, may I have your seat?' A moment of stark anomic panic overcame me. But the man got right up and gave me the seat. A second blow was yet to come. Taking the man's seat, I was overwhelmed by the need to behave in a way that would justify my request. My head sank between my knees, and I could feel my face blanching. I was not role-playing. *I actually felt as if I were going to perish.* (quoted in Blass 2004, 174)

Milgram's experience illustrates a familiar point. It is emotionally taxing to violate social and moral rules. In this discussion, I want to explore the connection between emotion and moral judgment and offer a diagnosis of Milgram's misery.

ISSN 1386-9795 print/1741-5918 online/06/010029-15

DOI: 10.1080/13869790500492466

This project has important implications for our understanding of practical reasoning. It bears, most directly, on the debate between motivational internalism and externalism in ethics. Can we make moral judgments without being motivated to act? If emotions are linked to moral judgments in an intimate way, then the answer may be negative. Moral judgments are intrinsically action-guiding. I will defend a version of the internalist position. But I will also identify ways in which such judgments can occur without placing immediate demands on behavior, and, in that regard, I will reconcile internalist moral psychology with some of the intuitions that drive externalism. I will suggest that there are different ways of conceptualizing obligations, which differ in their practical consequences. These differences bear on the nature of practical reasoning more generally.

Evidence for a Link Between Emotion and Moral Judgment

Philosophers have generally tried to establish the link between emotion and moral judgment by armchair reflection. I think philosophical analysis is a good way to make progress on the conceptual question: can one possess a moral concept without having certain sentiments? But conceptual questions are thorny, because many of our concepts are graded, open textured, or polysemous, and philosophical intuitions are, correlatively, divided. As a starting place, I want to focus on a more tractable question to consider: do our *ordinary* moral concepts (the ones we deploy in token thoughts most frequently) have an emotional component? This is essentially an empirical question. It's a question about what goes on in our heads when we use moral terms like 'good' and 'bad' or 'right' and 'wrong'. Empirical questions can be addressed using philosophical methods (philosophical intuitions can be treated as data), but laboratory studies are useful as well. In a spirit of methodological promiscuity, I propose to intermingle empirical and philosophical results.

Current evidence favors the conclusion that ordinary moral judgments are emotional in nature. I will present this evidence by defending a series of increasingly strong theses about how emotions and moral judgments interrelate. The first thesis that I want to defend is that emotions co-occur with moral judgments. This should not be terribly controversial. It is fairly obvious from experience that when we judge that a moral rule has been violated, we typically have a negative emotional response. This piece of introspective psychology has been confirmed again and again, in every study of what goes on in the brain during moral judgment.

For example, Moll, de Oliveira-Souza, and Eslinger (2003) measured brain activity as subjects evaluated moral sentences such as, 'You should break the law when necessary' in contrast with factual sentences such as, 'Stones are made of water'. In both cases, subjects simply had to answer 'right' or 'wrong'. They found that when subjects made moral judgments, as opposed to factual judgments, areas of the brain that are associated with emotional response were active. In a different study, Sanfey et al. (2003) measured brain activity as subjects played an ultimatum game. In each case, one player was asked to divide a monetary sum with another player. When the division was deemed inequitable, the second player had brain activity in areas associated with emotion. Berthoz et al. (2002) found similar engagement of emotion brain areas when subjects considered violations of social rules. For example, subjects were given a story about a dinner guest who, after tasting the food, rudely spat it out into a napkin without apology. Likewise, Greene et al. (2001) found emotion activation as subjects considered moral dilemmas, and

Kaplan, Freedman, and Iacoboni (forthcoming) found emotion activation as subjects looked at pictures of politicians who they oppose.

None of these findings is surprising. The brain scans simply add empirical support to a pretheoretical intuition that emotions arise when we respond to a wide range of morally significant events, including rudeness, unfairness, law-breaking, and saving lives. What neuroscience cannot at this stage establish is the specific role that emotions play. Are they mere effects of moral judgments or are they more intimately involved? For this question, we need other sources of evidence.

The second thesis I want to defend is that emotions influence moral judgments. A negative emotion can lead us to make a more negative moral appraisal than we would otherwise have. To prove this, Schnall, Haidt, and Clore (forthcoming) gave subjects a questionnaire with a series of vignettes and asked them to rate the wrongness of the actions described. For example, subjects read that:

> Frank's dog was killed by a car in front of his house. So he cut up the body and cooked it and ate it for dinner. How wrong was that?

Half the subjects who read these vignettes are seated at a nice clean desk. The other half are seated at a filthy desk, with a crusty drink cup, a chewed pencil, a used tissue, and a greasy pizza box. Subjects at the disgusting desk rated the vignettes as more wrong than subjects at the clean desk.

These findings are still open to a challenge. Perhaps negative emotions merely draw our attention to morally relevant features of a situation. It would be nice to establish that negative emotions can be sufficient for making negative moral judgments even when we have no other reason to think that a situation is wrong. Haidt and his collaborators have obtained evidence in support of this stronger sufficiency thesis. In one study, Wheatley and Haidt (forthcoming) hypnotized subjects to feel a pang of disgust when they heard the emotionally neutral word 'often'. They then presented these subjects with vignettes that either contained the word 'often' or a synonym. Some of these scenarios describe morally reprehensible characters, but others describe characters who are morally admirable. Subjects who are hypnotized to feel disgust when they hear the word 'often' judge that the morally admirable characters are morally wrong when that word appears in the vignettes! This suggests that a negative feeling can give rise to a negative moral appraisal without any specific belief about some property in virtue of which something is wrong. Similarly, Murphy, Haidt, and Björkland (forthcoming) asked subjects to justify their belief that a case of consensual incest between siblings is wrong. For every justification subjects provided, he gave a reply that rendered the justification irrelevant to the case. For example, most subjects claimed that if the siblings had sex they might have offspring with birth defects. Murphy replied by saying that the siblings used birth control. After several epicycles like this, a few subjects said that incest might be okay under these special circumstances, but the majority insisted that incest is still wrong in such cases, simply because it is disgusting.

Such findings suggest that we can form the belief that something is morally wrong by simply having a negative emotion directed towards it. In this sense, emotions are sufficient for moral appraisal. But are they necessary? I think there is evidence supporting a necessity thesis as well. In particular, I think emotions are needed for moral development. Unlike language, children need a lot of training to conform to moral rules, and parents spend a lot of time giving their children moral instruction. Interestingly, the three main techniques

that parents use to convey moral rules all recruit emotions (Hoffman 1983). One technique is power assertion (physical punishment or threat of punishment), which elicits fear. Another technique is called induction, which elicits distress by orienting a child to some harm she has caused in another person ('Look, you made your little brother cry!'). The third technique is love withdrawal, which elicits sadness through social ostracism ('If you behave like that, I'm not going to play with you!'). Each technique conditions the child to experience negative emotions in conjunction with misdeeds. This does not prove that emotions are necessary for moral development, but it is suggestive.

Stronger evidence for the necessity of emotions in moral development comes from research on psychopaths. Psychopaths are the perfect test case for the necessity thesis, because they are profoundly deficient in negative emotions, especially fear and sadness. They rarely experience these emotions, and they have remarkable difficulty even recognizing them in facial expressions and speech sounds (Blair et al. 2001, 2002). Psychopaths are not amenable to fear conditioning, they experience pain less intensely than normal subjects, and they are not disturbed by photographs that cause distress in us (Blair et al. 1997). This suggests psychopathy results from a low-level deficit in negative emotions. Without core negative emotions, they cannot acquire empathetic distress, remorse, or guilt. These emotional deficits seem to be the root cause in their patterns of antisocial behavior. I think that psychopaths behave badly because they cannot make genuine moral judgments. They give lip-service to understanding morality, but there is good reason to think that they do not have moral concepts—or at least they do not have moral concepts that are like the ones that normal people possess. Psychopaths acknowledge that their criminal acts are 'wrong' but they do not understand the import of this word. In a classic study of psychopaths, Cleckley (1941) compares psychopathy to colorblindness. He says a psychopath can say he understands good and evil, 'but there is no way for him to realize that he does not understand'. Blair (1995) investigated moral concepts in psychopaths more directly, and he found that they treat moral wrongs as if they were merely conventional. Psychopaths treat the word 'wrong' as if it simply meant 'prohibited by local authorities'.

Research on psychopathy suggests that emotions are developmentally necessary for acquiring the capacity to make moral judgments. The final thesis I want to advance is that emotions are also necessary in a synchronic sense. Here, some caution is needed. Obviously, we can say things like, 'killing is wrong' without feeling any emotion. We have committed these rules to memory. It's a bit like reporting that bananas are yellow without forming a mental image of yellowness. The necessity thesis I have in mind is dispositional. Can one sincerely attest that killing is morally wrong without being disposed to have negative emotions towards killing? My intuition here is that such a person would be confused or insincere. To support this intuition, we might imagine a person who knows everything non-emotional about killing. She knows that killing diminishes utility and that killing would be practically irrational if we universalized the maxim, thou shalt kill. Would we say of this person that she believes killing is wrong? It seems not. She could believe all these things without having any view about the morality of killing or even any comprehension of what it would mean to say that killing is wrong. Conversely, if a person did harbor a strong negative sentiment towards killing, we would say that she believes killing to be morally wrong, even if she did not have any explicit belief about whether killing diminished utility or led to contradictions in the will. These intuitions suggest that emotions are both necessary and sufficient for moral judgment. Empirical tests for the necessity of emotions

are harder to come by, but future experiments might examine this question. If we could hypnotically block negative emotions or induce strong positive affect and then ask subjects to make moral appraisals, we might find that negative appraisals become attenuated, especially when subjects are presented with cases than cannot be easily assessed using platitudes such as 'killing is wrong'. Moral blindness due to positive affect may explain why people who suffer from mania are often prone to antisocial behavior during manic episodes (American Psychiatric Association 1994, 330).

There is one final argument that I'd like to mention, for the thesis that emotions are necessary for morals. It is an argument from the anthropological record. If moral judgments were based on something other than emotions—something like reason or observation—we would expect more moral convergence cross-culturally. Reason and observation lead to convergence over time. Cross-culturally there is staggering divergence in moral values (reviewed in Prinz, forthcoming). The Guhuku-Gama of New Guinea and other headhunters think it is okay to kill innocent people; the Greek citizens of Ptolemaic Egypt married their siblings at a rate of up to 30%; the Aztecs of Mexico and countless small-scale societies indulged in cannibalism; the Romans filled arenas to watch gladiators slaughter each other; Thonga men have sex with their daughters before hunting; the women of China endured excruciating pain by binding their feet; gender inequity and slavery have been widely accepted, and widely condemned. Closer to home, we find interminable debates between liberals and conservatives. We also find regional differences: Southern white men are much more likely than their Northern counterparts to morally approve of violent reprisals for public insults, and other nonviolent offenses. These examples are not exotic. Any two randomly chosen cultures will have dramatic differences in moral values, and many of these differences (such as polygymy versus monogamy or Southern bellicosity versus Northern diplomacy) have no basis in different factual beliefs. This suggests that basic moral values do not have a purely cognitive source. Moral divergence does not directly demonstrate that emotions are a necessary component of morality, but it provides indirect evidence. If moral values are not driven by reason or observation, then it is plausible to think they hinge on culturally inculcated passions.

Sentimentalism

A Sentimentalist Theory of Moral Judgment

None of the empirical evidence that I have been discussing provides a demonstrative argument for any theory of moral judgment. Moral judgments might be correlated with and causally related to emotional responses without involving emotional responses essentially. The considerations presented on behalf of the claim that emotions are necessary for moral judgments point towards a very strong connection, but those considerations were far from decisive. At this stage, however, that may be enough. I want to advance a theory of moral judgment that systematizes the data I have been discussing. Other theories could be imagined. I want to suggest only that the theory I have in mind offers *an* explanation and perhaps a better explanation than many other accounts. The theory I have in mind is not new. It's a variant of an old theme, associated with the British moralists, especially Hume. Simply put, the theory says:

> To believe that something is morally wrong (right) is to have a sentiment of disapprobation (approbation) towards it.

This formulation needs to be refined in many ways, but it offers a helpful first approximation. As I will use the term, a sentiment is a disposition to have emotions. If you love chocolate, you will feel delighted when you see chocolate cake on the menu, and you will feel disappointed if the waiter then reports that they have run out. Sentiments of approbation and disapprobation are, likewise, constituted by different emotions on different occasions. For simplicity, I will not discuss approbation. Disapprobation encompasses emotions of blame, but there are a number of different emotions in this category. Which emotion we experience will depend on who is being blamed and for what. If I do something wrong, I may experience shame or guilt. If you do something wrong, I may experience anger, contempt, or disgust. There is very good evidence that different kinds of transgressions elicit different negative emotions. Shweder et al. (1997) have argued on the basis of anthropological evidence that there are three broad categories of moral rules. There are rules designed to protect persons (prohibitions against physical harm and rights violations), there are rules designed to protect the community (usually pertaining to rank or public goods), and rules pertaining to the perceived natural order (such as sexual mores or religious dietary rules). It turns out that these rules are associated with different emotions (Rozin et al. 1999). Crimes against persons elicit anger, crimes against community elicit contempt, and crimes against nature elicit disgust. In addition, the intensity of the felt emotion can vary with the wrongness of the action. The range of disapprobation emotions may be extended by distinguishing subtypes in each emotion category. One dimension of variation is intensity. Consider anger. If you harm someone badly, I may experience fury, but if the harm is petty, I may just shake my head with vague annoyance. We also distinguish subtypes of anger as a function of the eliciting conditions. Anger is labeled indignation when elicited by injustice, and rage when elicited by a physical assault. Anger may also change its character as a function of our relation to the transgressor. If a friend mistreats us, we may experience sullen brooding, and if a stranger mistreats us we may experience wrath. A full account of disapprobation should spell out all of these variations.

The sentimentalist thesis asserts that, when we judge that something is wrong, one or another of these emotions will ordinarily occur, and that the judgment will be an expression of the underlying emotional disposition. A standing judgment that something is wrong consists in the standing disposition (or its categorical basis), and an occurrent judgment will ordinarily contain a specific emotion that manifests the disposition. The emotion serves as the vehicle of the concept 'wrong' in much the same way that an image of some specific hue might serve as the vehicle for the thought that cherries are red. Tokens of the concept 'wrong' may be identical to emotions, but we can have and self-ascribe standing beliefs about wrongness without any emotions—a topic I will return to below.

When I say that moral judgments express sentiments, I do not mean to imply that moral judgments are merely expressive. I am not endorsing expressivism here. I prefer sensibility theories, according to which moral concepts refer to response-dependent properties (see Dreier 1990; Johnston 1989; McDowell 1985; McNaughton 1988; Prinz forthcoming; Wiggins 1991; Wright 1992). Moral judgments express sentiments, and sentiments refer to the property of causing certain reactions in us. The reactions in question are emotions, which I regard as feelings of patterned bodily changes (Prinz 2004). Sentiments often refer to response-dependent properties. If I say chocolate is likeable, I ascribe to chocolate the property of causing, e.g., pleasure in me. If I say Buster Keaton is funny, I ascribe to him the property of causing amusement in me. I may not realize that these are the properties I am ascribing. We sometimes assume that likeability or humorousness is an intrinsic

property, in just the way we assume that blue is a feature of surfaces, not a power that surfaces have to cause experiences in us. We project our experiences onto the world. I think our sentimental concepts are neutral with respect to what kinds of properties they ascribe. They do not specify one way or the other whether we are referring to something intrinsic or relational. We find it easy to imagine that moral properties inhere in the world, but not incoherent to suppose that they depend on reactions in us. (Compare the concept 'delicious'.)

Sentimentalism so-defined has a major advantage over expressivism. Moral judgments are truth-apt, if they refer to response-dependent properties, just as their surface form would suggest. But this immediately raises a rather embarrassing question. If 'wrong' refers to a response-dependent property, whose responses matter? And under what conditions? Here, I think something like Dreier's speaker-relativism is right. When I say that something is wrong, I refer (perhaps unwittingly) to the property of causing emotions of blame in me. (Or perhaps, saying that something is wrong means that it causes emotions of blame in us, where the 'us' refers to a group of people to whom I would morally defer. I leave this complexity out in what follows.) Speaker-relativism raises two immediate problems.

The first problem has to do with error. If 'wrong' referred to whatever causes disapprobation in me, then I could not judge something to be wrong in error. To avoid this consequence, we must idealize. We should say that the word 'wrong' refers only to those things that irk me under conditions of full factual knowledge and reflection, and freedom from emotional biases that I myself would deem as unrelated to the matter at hand. Neo-sentimentalists sometimes make this point by making a metacognitive move: wrong is not just that towards which I have a sentiment of disapprobation, it is that which I take to warrant such a sentiment. I think this metacognitive move is problematic and unnecessary. It is problematic because it requires that every moralizer have concepts of sentiments in addition to the sentiments themselves; this is not true of children (Nichols 2004). And it also requires that we have a concept of moral warrant, as opposed to other forms of warrant, and that risks the introduction of circularity in the account of moral judgment (D'Arms and Jacobson 2000). Fortunately, the metacognitive move is unnecessary. There are numerous theories of error that do not depend on metacognition in the literature on psychosemantics (e.g., Dretske 1988; Fodor 1990). Any one of these might be applicable here (e.g., knee-jerk sentimental reactions may be asymmetrically dependent on reflective reactions). There is also a more straightforward solution available, once we draw a distinction between emotions (which are occurrent states), and sentiments, which are dispositions to have emotions. Basic moral values may consist in having sentiments associatively linked in long-term memory to specific kinds of actions, abstractly construed. We might have a negative sentiment towards betrayal. Some action might cause us to have an emotion of blame because we mistake it for an instance of betrayal, even though it is not. This would be an error. On this analysis, we do not need to have metacognitive policies concerning our sentiments (though we might); rather, we have sentimental policies concerning types of action. A judgment that some action is wrong counts as erroneous if that action is not an instance of a type towards which we have a sentimental policy.

The second problem raised by speaker-relative response-dependent theories of moral concepts has to do with disagreement. If 'wrong' means 'wrong for me' then debates about what is wrong turn out to be spurious. You say infanticide is wrong, and I say it's not wrong, but we are really using the term in different ways. That's an unhappy result. There are

several familiar things to say in response. First, if we have the same basic moral values, then wrong-for-me and wrong-for-you may be extensionally equivalent; our debate may turn on factual differences about whether newborns have certain capacities. Reasoning is integral to moral judgment precisely because reasoning is often necessary to determine whether a particular form of conduct is an instance of some more general action type towards which we already have a moral sentiment. Second, we have good reason to debate because morality has practical consequences, and no policy can both allow and prohibit infanticide. We may be foolish to think we can resolve the debate rationally, but, in human life, even factual debates are often driven by rhetoric. Third, we may have some shared basic values which can be used to find a common denominator. Fourth, there is a good explanation for why we think our debate is legitimate: we tend to project sentimental properties onto the world (deliciousness, likeability, funniness) without realizing they are response-dependent. Fifth, the claim that moral debates are spurious explains why so many seem to be interminable; people reside in different moral worlds.

Obviously, much more would need to be said to defend this version of sentimentalism. I cannot offer a full account, much less a full defense here. For present purposes, there is no need to dwell on the details. I am content with the conclusion that to harbor a moral belief is to have a sentiment of approbation or disapprobation. I think that hypothesis makes the most sense of the data adduced in the previous section. Emotions co-occur with moral judgments, influence moral judgments, are sufficient for moral judgments, and are necessary for moral judgments, because moral judgments are constituted by emotional dispositions (either standing dispositions or manifest dispositions). At least our ordinary moral concepts seem to have this character (I will discuss some abnormal moral judgments below).

Explanatory Fruits

The sentimentalist theory of moral judgment explains the empirical data presented earlier, but that is not the end of its explanatory contribution. Sentimentalism can also explain three other things that are integral to our ordinary understanding of moral judgments.

First, sentimentalism explains the link between emotion and motivation. Philosophers disagree about whether moral judgments are intrinsically motivating, but they all admit that moral judgments characteristically give rise to motivational states. Moral judgments are closely associated with ought-judgments, and ought-judgments are characteristically action-guiding. Sentimentalism explains why there is such a rapid move from thinking an action is wrong to thinking I ought to prevent or avoid that action. If sentimentalism is true, thinking that an action is wrong disposes one to having negative emotions towards it, and negative emotions are inhibitory: they promote avoidance, ceasing, intervention, withdrawal, and, when anticipated, preventative measures. Beliefs about obligations are not add-ons to beliefs about wrongness; beliefs about wrongness carry the motivational force that we experience as being under an obligation. Consequently, moral judgments vie for control of the will. When they occur, we are thereby motivated to act.

Second, sentimentalism offers an attractive explanation of the distinction between moral and conventional rules (compare Blair 1995, Nichols 2004). By the time children are 3, they recognize that some rules are moral (e.g., don't hit other children) and others are merely conventional (don't talk without raising your hand). Children and adults consider moral transgressions more serious and less dependent on authorities. If the teacher says it's

okay to talk without raising your hand, then that's fine. If the teacher says it's okay to hit the child next to you, that does not make it okay. How do we draw this distinction? Why is it available to such young children? The answer may be that moral rules are directly grounded in the emotions. When we think about hitting, it makes us feel bad, and we cannot simply turn that feeling off. Hitting seems phenomenologically wrong regardless of what authorities say. We are less emotional about conventional rules. Speaking without raising your hand is bad, but it does not elicit rage or guilt. It is more likely to elicit embarrassment, but embarrassment is keyed to the reactions of others, so it diminishes when we engage in behavior that is accepted by those around us. Parents are much more likely to use emotional conditioning when teaching moral rules, and both children and adults are much more likely to mention emotions when justifying these rules. This explanation of the moral/conventional distinction predicts that we will tend to moralize mere conventions if we learn them through a process of emotional conditioning. Thus, a person who is emotionally conditioned to obey certain religious dietary rules may tend to treat these rules moralistically even if she recognizes that they are mere conventions. Likewise, conventional rules of etiquette that naturally elicit negative emotions when violated (e.g., spitting in public) may be more likely to be treated moralistically than other rules of etiquette (Nichols 2002).

There is a third explanatory fruit of sentimentalism. Early sentimentalists in Britain often found themselves in debates with intuitionists. Intuitionists believe that moral judgments are self-justifying; they do not stand in need of independent argumentative support. In this respect moral judgments are like certain perceptual judgments or mathematical judgments. It is difficult to come up with arguments for the self-justification thesis, and, indeed some intuitions simply assert that it is obviously true. They seem to base this assertion on the phenomenology of moral judgments: moral judgments *seem* self-evident. I think sentimentalism can explain this phenomenology. Sentimental judgments generally seem self-evident. It is evident to me that Buster Keaton is funny, because he makes me laugh. It is evident to me that chocolate is delicious because it induces pleasure when I taste it. It would be somewhat perverse to demand more evidence than this. Likewise, emotionally grounded moral judgments have a kind of perception-like immediacy that does not seem to require further support. We can feel that killing is wrong. Indeed, far from opposing intuitionism, sentimentalism offers one of the most promising lines of defense. The judgment that something is funny is justified by our amusement, because causing amusement is constitutive of being funny. If moral judgments are sentimental, and they refer to response-dependent properties, then the judgment that killing is wrong is self-justifying because killing elicits the negative sentiment expressed by that judgment and having the power to elicit such negative sentiments is constitutive of being wrong. Sentimentalism explains the phenomenology driving intuitionism, and it shows how intuitionism might be true.

Dispassionate Moralizing

Parasitic Dispassionate Moral Judgments

If sentimentalism is right, then our ordinary moral judgments are bound up with emotions. More specifically, the moral terms 'right' and 'wrong' express sentiments, and the token judgment that something is right or wrong may express an occurrent manifestation of a sentiment, which is an emotion. If you cross me, I may experience anger, and

that anger qualifies as a token of my concept WRONG. It is the experience of that anger that alerts me to the fact that you have done something morally questionable, and I give voice to the anger when I judge that you were wrong to treat me that way. For sentimentalists, ordinary moral judgments are hot.

This consequence is consistent with the empirical evidence. We do usually experience emotions when making moral appraisals. But the empirical evidence tells us how things ordinarily are, not how they must be. Brain scans do not carve up modal space. Surely there are conditions under which we make moral judgments dispassionately.

The theory that I have been defending allows this. A sentiment is an emotional disposition, and I can have a sentiment without manifesting it. I love John Coltrane, and I can truly self-ascribe this sentiment, but I am not always experiencing the hedonic rush that I get when I listen to *A Love Supreme*. Likewise, I can testify that I think gender discrimination is wrong without experiencing any outrage. In both cases, however, experienced emotions serve as a sincerity condition. If Coltrane never thrills me, then I am being disingenuous when I claim to be a fan. Likewise, if I am never outraged by gender discrimination, I am paying lip service to equity.

Sentimentalism also allows that one can *ascribe* moral judgments dispassionately. This occurs when discussing the attitudes of members of other groups. Anthropologists describe the moral values of head-hunters without adopting those values. They can even derive those values from purely factual information. If they observe that head-hunters celebrate the killing of innocent people, they can conclude that head-hunters find such murders morally commendable; they believe people ought to take heads. In this way, we can derive a moral rule from descriptive facts. But notice that the anthropologist can conclude only that head-hunters are under an obligation, given head-hunting morality; they cannot conclude that head-hunters 'ought' to take heads. When we utter 'ought' we express our own sentiments, and factual knowledge is not sufficient for having sentiments. So we can derive an obligation from an 'is' but not an ought from an 'is'. We can make dispassionate judgments about morals, but not dispassionate moral judgments.

Against this claim, moral externalists argue that there is conceptual space for people who make moral judgments without being in the least bit moved. They call such knavish individuals amoralists. In the real world, psychopaths are as close as we can find to amoralists: when they say that killing is wrong, they have no inclination to refrain from killing. But I think psychopaths are like anthropologists. They report on morality without making moral judgments. Indeed, they do not even get our morality right; they fail to distinguish moral and conventional rules. The concepts that psychopaths express when they use the words 'right' and 'wrong' differ from our concepts in both sense and reference. They can mention these concepts, but they can't use them.

In sum, sentimentalism allows two kinds of dispassionate judgments about morality. First, we make judgments that express sentiments, even when those sentiments are not at the present time manifesting themselves as occurrent emotions. Such judgments are not immediately motivating, but they express a dispositional state that would motivate under the right circumstances. Second, there are judgments that refer to the motivational states of others, as when we talk anthropologically about the obligations of other cultural groups. Psychopaths may use moral talk in this way. In both of these cases, the dispassionate moral judgment is parasitic on the passions.

Of course, externalists will insist that there is a way to make moral judgments that are neither sentimental nor parasitic on sentimental judgments. Call these external moral

judgments. Simply stipulating that external moral judgments are possible will not advance debate. Internalists don't share that intuition. I have argued that there is empirical evidence for a link between emotions and moral judgments, and apparent cases of dispassionate moral judgments are parasitic on passionate cases. Judgments that did not have a sentimental component would differ from the ordinary cases of moral judgments, and I see little reason why we should call such judgments moral. This is a case where we can move beyond the typical intuition mongering in philosophy and use empirical findings to help adjudicate an otherwise interminable philosophical debate. My assessment is that internalists come out ahead (though see Kennett 2002, for a different assessment of the empirical evidence).

Can There Be Non-parasitic Dispassionate Moral Judgments?

I can think of only one plausible strategy for defending the claim that there are external moral judgments, and I think that strategy will not succeed. It goes like this. I have argued that moral judgments are *ordinarily* sentimental in both form and content. Ordinary moral judgments are constituted by sentiments, and they represent the response-dependent property of causing sentimental responses in us. Someone might agree with the first half of this, and deny the second. Someone might argue that moral judgments are ordinarily, as a matter of fact, sentimental in form, but they do not refer to response-dependent properties. Instead, they refer to something else, to which our sentiments happen to be well attuned. For example, the bad might be that which fails to maximize utility or that which, when universalized, leads to a contradiction in the will. In other words, one could adopt a Humean view of the *sense* of ordinary moral concepts, while adopting a Millian or Kantian view of the *reference* of ordinary moral concepts. Perhaps our sentiments designate Millian or Kantian properties.

If this hybrid theory of moral concepts were right, then, in principle, one could have a concept that was co-extensive with ordinary moral concepts but utterly dispassionate. The concept FAILS TO MAXIMIZE UTILITY would be co-extensive, on one version, with the sentiment of disapprobation. On this view, the judgment that murder is wrong might ordinarily be affective but, by substitution of co-referring concepts, the sentimental version of the concept WRONG could be swapped out for the affect-neutral concept FAILS TO MAXIMIZE UTILITY. This would be the same judgment, in some sense. It would, at least, be a judgment with the same truth conditions. Consequently, we could call it a moral judgment. If sentiments refer to something other than response-dependent properties, then dispassionate moral judgments are possible. Such dispassionate moral judgments might be derived, as in this example, by substitution of co-referring concepts, but they would not be parasitic on the sentiments. If we encountered a Vulcan species whose members had no emotions, we might credit them with moral judgments simply in virtue of the fact that they could make judgments that have the same truth conditions as the judgments we make with our sentiments.

This strategy for defending dispassionate moral judgments is doomed to fail, I believe. It requires the highly dubious premise that our ordinary moral concepts—the ones constituted by our sentiments—might refer to the kinds of response-independent properties that have been celebrated by normative ethicists. But this is vanishingly unlikely. No plausible theory of reference could possibly deliver this result. There are effectively two ways for a concept to refer: by description or by causation. The problem is that moral

concepts are not descriptively or causally linked to the kinds of properties that normative ethicists like to talk about.

Let's begin with descriptive theories of reference. If sentimental concepts refer to Kantian or Millian properties by description, then we should be able to figure that out by conceptual analysis. It should be a conceptual truth that the good is that which maximizes utility, say, and a contradiction to suppose that there can be a good course of action that fails to maximize utility. But that is manifestly not a conceptual truth. If there is any lesson we can extract from Moore's vexed open question argument, it is that there is no analytic tie between ordinary moral concepts and the descriptive concepts that designate the properties implicated in the moral theories of Mill, Kant, and other normative ethicists. It is an open question whether the good is that which maximizes utility, so utility maximization cannot be the descriptive content of 'good'. Contrast this with the case of paradigmatic descriptive concepts. It is not an open question whether bachelors are unmarried or whether uncles are brothers.

One might try to defend the claim that moral concepts refer descriptively to Kantian or Millian properties by suggesting that their descriptive contents are not known consciously. Ever since Plato, philosophers have believed that some concepts have descriptive content that requires agonizing philosophical toil to bring out. Perhaps we are implicitly committed to the view that the good is that which maximizes utility, say, and we just don't realize it. After a good dose of philosophical midwifery, we might realize this, and the apparently open question will close.

I think this is wishful thinking. Normative ethical theories are certainly seductive, and the best of them may indeed uncover genuine demands on action. Perhaps we are rationally obligated to be Kantians or Millians. I don't want to suggest that these normative theories are false. Instead, I want to question their descriptive accuracy. I don't think they can be defended as plausible analyses of ordinary moral concepts. These and other normative theories are best construed as correctives. They are best understood as proposals for replacing ordinary moral concepts, not as analyses of them. One reason for this diagnosis is that normative ethical theories tend to promote impartiality, and ordinary moral concepts tend to be partial. This is obvious when we look at moral values cross-culturally. The headhunting Guhuku-Gama of New Guinea, for example, think that is morally wrong to kill a member of their kin group, and perfectly fine to kill others. This is not an inconsistent position: they think it is morally acceptable for others to kill their kin. Moral considerability is a function of connectedness to the moral agent. Likewise, when the great Chinese philosopher Mo Tsu began to advocate universal love in the fifth century B.C.E., Mencius, speaking for the Confucian mainstream, complained that this would 'reduce humans to the level of beasts' (Harris 1989, 455). Pluralization of culture has probably led to an increased tendency to morally praise impartiality, but few ordinary moralizers would go so far as Mill or Kant. For example, the overwhelming majority of Americans oppose 'outsourcing' jobs, and few spend any resources combating world hunger. Most would probably agree that they have strong moral obligations to friends and kin, and few show signs of feeling morally obligated to strangers in developing nations. Helping strangers is charity, not responsibility. Obviously, much more would need to be said to prove that our moral concepts do not have descriptive contents that conform to the strictures of Kant, Mill, or other normative ethicists. The point here is that there is little *prima facie* evidence for such conformity.

If ordinary moral concepts do not refer to Kantian or Millian properties by description, then perhaps they refer to such properties by causation. This is the only serious option left

for those who want to argue that ordinarily moral concepts have contents that can be characterized in a way that does not make reference to our sentiments. To explore this possibility, let's consider the kind of causal theories that are most popular in contemporary semantics: let's assume that non-descriptive concepts refer to whatever reliably causes them to be tokened. To make good on the claim that ordinary moral concepts refer to Kantian or Millian properties in this way, we would have to show that our moral sentiments are reliably caused by Kantian or Millian properties. For example, we might try to show that the sentiment of disapprobation is reliably caused by failures to maximize utility. It should be immediately obvious that this is a silly suggestion. All too often, people morally applaud actions that fail to maximize utility. We may applaud actions that increase utility, but we do not insist on maximization. Most actions that increase utility actually fail to maximize, and should therefore trigger disapprobation, if moral sentiments were tuned to Millian properties. But maximization failures have no special causal link to disapprobation. Likewise, we have no strong disposition to condemn actions that fail Kantian tests. Every time we put on an article of clothing, eat a bit of food, play a song on the radio, or read a book, we are using other people as means, rather than ends. Every time we try to beat rush-hour traffic, we are doing something that cannot be willed as a universal law. None of these mundane actions promotes strong feelings of anger or guilt.

More generally, I don't think we can single out any unique property as the reliable cause of our moral sentiments. The range of things that trigger disapprobation is radically disunified. For example, we condemn murder, bestiality, destruction of nature, inequitable distributions, and plural marriages. In each society, people condemn a different range of things, and everything condemned by one society is applauded by another. There does not seem to be any common denominator here. There is no single response-independent property of actions or events that reliably triggers disapprobation in any of us. The only thing that unifies iniquities is the responses that they cause. It is for that reason that I think WRONG designates the power to cause disapprobation, and not some other property that can be characterized without reference to our responses.

I conclude that ordinary moral concepts do not refer to properties that can be coherently characterized without reference to our sentiments. Therefore, there is no concept that co-refers with our sentiments that does not either contain or advert to our sentiments. Any concept that is co-referential with our moral sentiments is parasitic on those sentiments. Normative ethicists introduce concepts that refer to something other than what our ordinary moral concepts refer to. Is there any reason to call the concepts that they introduce 'moral'? I think that is a question without a definite answer. Normative ethicists introduce concepts that, like ordinary moral concepts, are designed to regulate behavior, but these concepts are different from the concepts we ordinarily express when we use words such as 'right' and 'wrong'. Whether we call such concepts 'moral' or not is a matter of choice. The point I want to make is that the concepts we ordinarily express using moral vocabulary are linked essentially to our sentiments, and, in this sense, motivational internalism is true.

Conclusion

I have presented empirical evidence for a link between ordinary moral concepts and emotions. I argued that the empirical evidence is best explained by a sentimentalist theory of moral concepts, and that sentimentalist theory bears some philosophical fruit. It is a consequence of this theory that moral judgments are generally motivating because emotions

have motivational force. In some cases, moral judgments occur without emotions, but these are parasitic on the emotional cases. Normative judgments that are not parasitic on the emotions might be called 'moral', but not in the ordinary sense of the term.

ACKNOWLEDGEMENTS

I am very grateful to Jeanette Kennett and Philip Gerrans for organizing the marvelous conference that led to this special issue. I also learned a great deal (not reflected here) from commentaries by Karen Jones and by Ruth Chang at another venue. Various portions of this material were presented to audiences at Birkbeck, Brown, Cincinnati, the CUNY Graduate Center, the Ecole Normale Superieure, Georgia Tech, Leeds, Monash, Minnesota, Northwestern, Stockholm, Texas Tech, and Toronto. I am grateful to audiences in all those places. I also owe Melissa van Amerongen for catching a number of errors in the manuscript.

REFERENCES

AMERICAN PSYCHIATRIC ASSOCIATION. 1994. *Diagnostic and statistical manual of mental disorders*. 4th ed. (DSM-IV). Washington, D.C.: American Psychiatric Association.

BERTHOZ, S., J. L. ARMONY, R. J. R. BLAIR, and R. J. DOLAN. 2002. An fMRI study of intentional and unintentional (embarrassing) violations of social norms. *Brain* 125: 1696–708.

BLAIR, R. J. R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1–29.

BLAIR, R. J. R., E. COLLEDGE, L. MURRAY, and D. G. MITCHELL. 2001. A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology* 29: 491–98.

BLAIR, R. J. R., L. JONES, F. CLARK, and M. SMITH. 1997. The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology* 34: 192–98.

BLAIR, R. J. R., D. G. V. MITCHELL, R. A. RICHELL, S. KELLY, A. LEONARD, C. NEWMAN, and S. K. SCOTT. 2002. Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals. *Journal of Abnormal Psychology* 111: 682–86.

BLASS, T. 2004. *The man who shocked the world: The life and legacy of Stanley Milgram*. New York: Basic Books.

CLECKLEY, H. M. 1941. *The mask of sanity: An attempt to reinterpret the so-called psychopathic personality*. St Louis, Mo.: The C. V. Mosby Company.

D'ARMS, J., and D. JACOBSON. 2000. Sentiment and value. *Ethics* 110: 722–48.

DREIER, J. 1990. Internalism and speaker relativism. *Ethics* 101: 6–26.

DRETSKE, F. 1988. *Explaining behavior*. Cambridge, Mass.: MIT Press.

FODOR, J. A. 1990. *A theory of content and other essays*. Cambridge, Mass.: MIT Press.

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: 2105–8.

HARRIS, M. 1989. *Our kind: The evolution of human life and culture*. New York: Harper Perennial.

HOFFMAN, M. L. 1983. Affective and cognitive processes in moral internalization. In *Social cognition and social development: A sociocultural perspective*, edited by E. T. Higgins, D. N. Ruble, and W. W. Hartup. Cambridge: Cambridge University Press.

JOHNSTON, M. 1989. Dispositional theories of value. *Proceedings of the Aristotelian Society* 63 (Supplement): 139–74.

KAPLAN, J. T., J. FREEDMAN, and M. IACOBONI. Forthcoming. Us vs. them: Political attitudes and party affiliation influence neural response to faces of presidential candidates.

KENNETT, J. 2002. Autism, empathy and moral agency. *The Philosophical Quarterly* 52: 340–57.

MCDOWELL, J. 1985. Values and secondary qualities. In *Morality and objectivity*, edited by T. Honderich. London: Routledge & Kegan Paul.

MCNAUGHTON, D. 1988. *Moral vision: An introduction to ethics*. Oxford: Blackwell.

MOLL, J., R. DE OLIVEIRRA-SOUZA, and P. J. ESLINGER. 2003. Morals and the human brain: A working model. *Neuroreport* 14: 299–305.

MURPHY, S., J. HAIDT, and F. BJÖRKLUND. Forthcoming. Moral dumbfounding: When intuition finds no reason.

NICHOLS, S. 2002. On the genealogy of norms: A case for the role of emotion in cultural evolution. *Philosophy of Science* 69: 234–55.

NICHOLS, S. 2004. *Sentimental rules: on the natural foundations of moral judgment*. Oxford: Oxford University Press.

——. Forthcoming. Sentimentalism naturalized. In *The psychology and biology of morality*, edited by W. Sinnott-Armstrong.

PRINZ, J. J. 2004. *Gut reactions: A perceptual theory of emotion*. New York: Oxford University Press.

——. Forthcoming. *The emotional construction of morals*. Oxford: Oxford University Press.

ROZIN, P., L. LOWERY, S. IMADA, and J. HAIDT. 1999. The CAD triad hypothesis: A mapping between three moral emotions (contempt, anger, disgust) and three moral codes (community, autonomy, divinity). *Journal of Personality and Social Psychology* 76: 574–86.

SANFEY, A. G., J. K. RILLING, J. A. ARONSON, L. E. NYSTROM, and J. D. COHEN. The neural basis of economic decision-making in the ultimatum game. *Science* 300: 1755–58.

SCHNALL, S., J. HAIDT, and G. CLORE. Forthcoming. Irrelevant disgust makes moral judgment more severe, for those who listen to their bodies.

SHWEDER, R. A., N. C. MUCH, M. MAHAPATRA, and L. PARK. 1997. The 'big three' of morality (autonomy, community, and divinity), and the 'big three' explanations of suffering. In *Morality and health*, edited by A. Brandt and P. Rozin. New York: Routledge.

WHEATLEY, T., and J. HAIDT. 2005. Hypnotically induced disgust makes moral judgments more severe. *Psychological Science* 16: 780–84.

WIGGINS, D. 1991. A sensible subjectivism. In *Needs, values, truth: Essays in the philosophy of value*. Oxford: Blackwell.

WRIGHT, C. 1992. *Truth and objectivity*. Cambridge, Mass.: Harvard University Press.

Jesse Prinz, Department of Philosophy, Campus Box 3125, University of North Carolina, Chapel Hill, NC 27599, USA. E-mail: jesse@subcortex.com