

## *What neuroscience can (and cannot) contribute to metaethics*

Richard Joyce

Penultimate draft of the article appearing in W. Sinnott-Armstrong (ed.), *Moral Psychology Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (MIT Press, 2008): 371-394.

Suppose there are two people having a moral disagreement about, say, abortion. They argue in a familiar way about whether fetuses have rights, whether a woman's right to autonomy over her body overrides the fetus's welfare, and so on. But then suppose one of the people says "Oh, it's all just a matter of opinion; there's no *objective fact* about whether fetuses have rights. When we say that something is morally forbidden, all we're really doing is expressing our disapproval of it." The other person protests: "No!—that's totally wrong. Of course there are objective moral truths." And suppose their dispute now settles on this new matter of whether there are objective moral facts, and the debate continues. They've now stopped discussing abortion, and are discussing the nature of any moral debate about abortion; they've stopped doing ethics, and started doing metaethics. If wondering about what one morally ought to do is to engage in ethical thought, then wondering about what one is doing when one wonders about what one morally ought to do is to engage in metaethical thought. When we make public moral judgments are we stating facts, or are we just expressing our opinions? And if there are moral facts, then what kind of fact are they? Can moral judgments be true or false? Can moral judgments be justified, and if so how? These kinds of questions are the domain of metaethics. (I'm not saying that there is a crisp and principled line to be drawn between meta-ethics and regular ethical discourse; but it is, if nothing else, a pedagogically useful division of labor.)

Thus the issue of whether a body of empirical data can have any metaethical implications is different from the issue of whether it can have *ethical* implications. We might be wary of the claim that purely descriptive information could have any ethical implications for the oft-cited reason that one cannot derive an "ought" from an "is." (As a matter of fact, the ban on deriving an "ought" from an "is" is a piece of philosophical dogma that I don't subscribe to, but let that pass for now.) Even if there were an *a priori* prohibition on deriving evaluative conclusions from factual premises, this need not stand in the way of *metaethical* implications being drawn from factual premises, for a metaethical claim is not an ethical "ought" claim; it is more likely to be a claim about how we use the word "ought" in ethical discourse—which is a perfectly empirical matter.<sup>1</sup> Indeed, I think metaethicists tend to be very open to the possibility of empirical work shedding light on their field. Recently there has appeared a wealth of data that a metaethicist might take an interest in: findings from neuroscience, from developmental psychology, evolutionary psychology, social psychology, evolutionary biology, experimental economics, cross-cultural anthropology, and even primatology. But although all this research is doubtlessly *of interest* to a metaethicist, does any of it actually contribute to the resolution of any of our perennial metaethical questions? This chapter confines itself to addressing just two points: *Can neuroscience support moral emotivism?* and

---

<sup>1</sup> In so far as some metaethicists offer prescriptions about how the word "ought" *ought* to be used, metaethics sometimes steps beyond the descriptive. Even in such cases, however, metaethicists are still not pushing *ethical* "ought" claims.

*Can neuroscience undermine moral rationalism?* (Since one of my main intentions is to disambiguate the diverse ways in which the terms “emotivism” and “rationalism” are used, it would be pointless to attempt a generic definition of either here at the outset.)

The first will not require much discussion to arrive at a negative answer. Perhaps this will be of no surprise to philosophical readers—indeed, many may be astonished that anyone would be tempted to answer the question positively. Yet there has crept into the literature a tendency to describe certain empirical data—both from neuroscience and from social psychology—as supporting “emotivism.” Jonathan Haidt leaves us in little doubt that he takes his research to vindicate “Hume’s emotivist approach to ethics” (2001: 816), and elsewhere he is described explicitly as an “emotivist” by Joshua Greene et al. (2004: 397). Greene and colleagues also take their own neuroscientific research to confirm a view that steers “a middle course between the traditional rationalism and the more recent emotivism that have dominated moral psychology” (2001: 2107). Haidt (2003: 865) describes Freud as an emotivist (or, at least, “a rare but ready ally” of emotivism) on the grounds that he took reasoning to be often just a rationalization of subconscious desires; and this terminology is taken up by others who assert that “emotivist perspectives on moral reasoning hold that emotional reactions precede propositional reasoning” (Fessler et al. 2003: 31). It should be fairly obvious to anyone familiar with the metaethical tradition that these empirical scientists are simply using the word “emotivism” differently from philosophers. It would be fruitless and churlish to make this the basis of criticism; psychologists may, of course, use the word how they wish. But it is nevertheless important to forestall any possible misconception that the empirical data support *what philosophers call* “emotivism,” and I am satisfied if the following comments achieve this modest aim.

The question of whether neuroscience can undermine moral rationalism suffers from the same cross-disciplinary terminological confusion, since in the mouths of many psychologists “emotivism” and “rationalism” are taken to denote two poles of a continuum, defined according to which faculty is in the driving seat in the production of moral judgment. Again: Haidt explicitly takes himself to be providing evidence against “rationalist” models of moral judgment, while Greene et al. also talk of their findings challenging “rationalism” (2004: 397). For metaethicists, by contrast, although emotivism and rationalism are generally considered contraries, they are very different *kinds* of theory: One is a theory about the linguistic function of moral utterances, the other is usually taken to concern the justificatory basis of actions. To a certain degree this is just another case of two academic disciplines using the same words in a divergent manner, and to that extent there is little to get excited about. But it is worrying that the empirical scientists in question often purport to connect their findings to the moral rationalism that is found in the *philosophical* tradition (e.g., they mention Kant, Rawls, etc.), and indeed there is some work that tries to use empirical findings to refute specific versions of rationalism that have been formulated and expounded by moral philosophers (e.g., Nichols 2004, to be discussed below). Most of this paper will be devoted to teasing apart different claims that might pass under the rubric “moral rationalism,” with an eye to gauging just what empirical research—especially that of a neuroscientific nature—might contribute. A sub-project of this investigation will be a brief discussion of whether evidence from neuroscience—specifically, the phenomenon of acquired sociopathy—helps break an impasse within metaethics over whether moral judgment necessarily implicates motivation.

## 1. Can neuroscience support moral emotivism?

Emotivism is a metaethical theory that had its heyday back in the 1930s and 1940s, but continues to attract a fair share of attention. Emotivism is a theory about moral language—most clearly construed as concerning the function of public moral utterances. It states that when we make a moral judgment we are not expressing a belief (i.e., are not making an assertion), but rather expressing some kind of conative mental state, such as a desire, emotion, or preference. Often moral emotivism is attributed to David Hume back in the 18th century, but I think this is quite mistaken. Hume certainly thinks that emotions (“passions”) play an important and possibly essential role in moral judgment, but one struggles to find him saying anything that implies that moral language *functions to express* emotions (see Joyce forthcoming *a*). One of the earliest clear statements of emotivism comes from A.J. Ayer in 1936, who famously claimed that the judgment “Stealing money is wrong” does not express a proposition that can be true or false, but rather it is as if one were to say “*Stealing money!!*” with the tone of voice indicating that a special feeling of disapproval is being expressed (Ayer [1936] 1971: 110). It was Ayer’s logical positivism that led him to such a view; convinced that all meaningful statements must be either analytic or empirically verifiable, and faced with a chunk of language that appears to be neither, the only way to grant the meaningfulness of this language was to deny that it consists of statements.

Emotivism can be presented either as a semantic or as a pragmatic theory. One might claim that the sentence “Stealing is wrong” *means* “Boo to stealing!” This is a semantic version since it makes a claim about what the sentence really means; it provides a translation scheme from something with a propositional structure (“Stealing is wrong”) to something non-propositional (“Boo to stealing!”); it claims that the moral predicates (“...is wrong,” etc.) are predicates only at the grammatical level, not at the logical level. Alternatively, moral emotivism can be presented as a pragmatic theory, holding that the meaning of moral judgments like “Stealing is wrong” is exactly what it appears to be (whatever that is), but that when we employ such sentences in moral discourse we are not asserting them—rather, we are by convention using them to express disapproval, or some other conative state. What these two types of emotivism have in common is that they claim that when we make a public moral judgment we are, despite appearances to the contrary, not expressing a belief. Let’s say, to make things simple, that they hold instead that we are expressing an emotion.

One might be tempted to think that a lot of recent research from psychology and neuroscience supports emotivism, since this research shows that emotions play a central role in moral deliberation. I leave it to others to present this evidence in detail (see Moll, Greene, Haidt, Nichols, Prinz: this volume). Looking no further than the abstracts of two central papers we find Moll and colleagues writing that “emotion plays a pivotal role in moral experience” (2001: 2730), while Greene and Haidt conclude that “recent evidence suggests that moral judgment is more a matter of emotion and affective intuition than deliberative reasoning” (2002: 517). Surely, one might think, if we find that when we hook up people’s brains to a neuroimaging device, get them to think about moral matters, and observe the presence of emotional activity, emotivism is supported?

No, it isn’t. We need to pay attention to what is meant by “express” when we talk about what kind of mental state a public utterance expresses. Sometimes “express” is used to denote a causal relation. If we say that by kicking over her brother’s sand castle Emily expressed her

anger, we may mean that anger caused her action or that it is an important element in an adequate explanation of the action. If it turns out that Emily in fact isn't angry at all, we will have to reject this explanation. But often "express" is used differently. When Emily later apologizes for kicking over the sandcastle, she expresses regret. Suppose, though, that Emily's apology is insincere, in the sense that she has not a glimmer of regret for what she did. This doesn't change the fact that she apologized. An insincere apology still succeeds in being an apology (just as an insincere promise is still a promise, and an insincere assertion is still an assertion). Nor does insincerity change the fact that Emily thereby expressed regret, for an apology *is* an expression of regret. Here "express" does not denote an explanatory or causal relation holding between Emily's utterances and her mental states; rather, it indicates a much more complex relation, holding between Emily, her brother, and a range of linguistic conventions according to which when a person utters "I'm sorry" in the appropriate circumstances then she has (among other things) *expressed regret*.<sup>2</sup> Thus it is perfectly possible that one can express regret over something when in fact one has no regret at all. This shows that the expression relation cannot be a causal or explanatory one, but is, rather, a matter of linguistic convention.

When the metaethical emotivist claims that moral judgments express emotions, he or she is using "express" in the same way as when we say that an apology expresses regret, or that an assertion expresses a belief. Once this is understood, it becomes apparent that the most that neuroscientific discoveries could establish is that public moral judgments are *accompanied* by emotions, and perhaps that they are *caused by* emotions—but further arguments would be needed to show that public moral judgments *express* those emotions. It is entirely possible that moral judgments are typically caused by emotional activity but nevertheless function linguistically as assertions (i.e., expressions of belief).

Hume, for example, favored a projectivist account of moral phenomenology; he spoke of the mind's "great propensity to spread itself on objects" (Hume [1740] 1978: 167), and claimed that "taste" (as opposed to reason) "has a productive faculty, and gilding and staining all natural objects with the colours, borrowed from internal sentiment, raises in a manner a new creation" (Hume [1751] 1983: 88). The idea is that certain events or states of affairs cause us to feel emotions—such as anger, disgust, or approval—which we then "project" onto our experience, seeing the world as containing qualities that it does not in fact contain. It's not that our emotions cause us to experience external events as having emotions (which would just be bizarre), but rather our emotions cause us to experience external events as having normative properties like *demanding anger*, or simply *being wrong*. It is often assumed that moral projectivism and moral emotivism go hand in hand (indeed, the terms are sometimes used as if synonyms), but this is mistaken. From the fact that its seeming to someone as if the situation has the property of being wrong is to be explained by the situation having prompted in him the emotion of anger or disgust, it does not follow that the way he articulates things—uttering "That situation is wrong"—*functions to express* those emotions. It is, in fact, hard to see how projectivism and emotivism even could go happily together. The crucial thing to notice is that projectivism implies an account of how the world *seems* to those who are doing the projecting: It seems to them as if it contains properties. Since we can assume that the

---

<sup>2</sup> "Among other things" indicates that although one cannot apologize without expressing regret, one can admit to having regret without thereby apologizing; an apology also requires an admission of responsibility, for example (Kort 1975; Joyce 1999).

language with which they discuss the matter will reflect their experience, then when they say things like “That act was wrong” it seems safe to assume, absent any reason to think otherwise, that they are expressing their belief that the external situation instantiates this property. But if they are expressing their beliefs on the matter (that is, *asserting* that the act is wrong) then they cannot simply be expressing their emotions.

My point here is not to argue in favor of moral projectivism, but simply use it to illustrate a possible scenario where emotions are centrally implicated in the production of moral judgments but nevertheless the public form of these judgments may be entirely assertoric. Nor am I claiming that projectivism entails the denial of emotivism, simply that moral cognitivism is its more natural partner.<sup>3</sup> The only way to settle the matter is to investigate directly the nature of the linguistic conventions surrounding our moral discourse, not the nature of neurological etiology. Uncovering such linguistic conventions may be in substantial part an empirical inquiry—we might think of it as “socio-linguistics.” To this extent I think that emotivism certainly could be supported by empirical evidence—but this evidence would concern *how we use moral language*. I take it that appealing to this kind of evidence—examining linguistic practices, uncovering shared intuitions about far-fetched thought experiments, etc.—is fairly pervasive in metaethics, and indeed in philosophy in general. But as to other forms of empirical evidence—in particular the evidence from neuroscience and psychology revealing what is going on in our brains when we make moral judgments—these findings, I think, should give no particular consolation to the moral emotivist.

## 2: Can neuroscience undermine moral rationalism?

We’ve just seen the dangers of thinking of emotivism simply as the view that morality necessarily has “something to do with” the emotions. By contrast, it is difficult to characterize *moral rationalism* more precisely than to say that it is the claim that morality necessarily has “something to do with” rationality. I say this because moral rationalists are a motley bunch. But it is precisely because of this indeterminacy that one cannot be too confident in saying that one argument, or one body of empirical evidence, “undermines rationalism.” I think that there are certain types of rationalism that do look shaky as the result of empirical evidence, but there are other kinds that remain untouched. And I’m also inclined to think that the latter kind of rationalism—the kind apparently immune from empirical debunking—is the more metaethically interesting variety. I’ll proceed by disambiguating three versions of moral rationalism: Psychological Rationalism, Conceptual Rationalism, and Justificatory Rationalism. Let us consider them in turn.

---

<sup>3</sup> Moral cognitivism is the view that public moral judgments typically express beliefs, i.e., are assertions. If one defines cognitivism as the view that moral judgments express *only* beliefs, and emotivism as the view that they express *only* conative states, then the two are obviously incompatible. But if these optional appearances of “only” are removed, then a mixed cognitivist/emotivist view becomes possible (and, indeed, I think there is much to be said in its favor). According to such a view, moral language is in this respect like certain pejorative terms: To call someone “a kraut,” for example, is both to express a belief (that the person is German) *and* to express an attitude of contempt. See Copp 2001; Joyce 2006a and forthcoming. Given the complicating possibility of this mixed view, what the footnoted sentence should really say is that the natural partner of projectivism is a metaethical view *that endorses a cognitivist element* (a comment that would, absent this accompanying explanation, be apt to confuse).

*Psychological Rationalism* is the view that moral decisions and moral deliberations causally flow from a “rational” faculty. The theory has a long and distinguished career in philosophy. Plato thought of moral judgments as the product of the rational faculty, which apprehends eternal moral truths. Aquinas thought that humans have an innate rational faculty called “synderesis” that informs us of our moral obligations. Hume’s well-known dichotomizing of the issue was focused on whether morals are “the product of reason” or “the product of passion.” To the extent that there is also a tradition of moral rationalism in psychology—represented by such figures as Piaget and Kohlberg—then Psychological Rationalism is its core thesis.<sup>4</sup>

We must begin by making some broad distinctions. First we should distinguish the claim that the activity of the rational faculty is necessary for moral judgment from the stronger claim that such activity is necessary *and sufficient*. According to strong Psychological Rationalism moral judgments are the product of the rational faculty *alone*. We should also distinguish synchronic from diachronic versions of Psychological Rationalism. The former holds that every moral judgment flows from activity in the agent’s rational faculty occurring at the time (or—fudging slightly—shortly before the time) of the moral judgment. A diachronic version allows that moral judgment may not be accompanied by such rational activity, but that such activity was necessary at some developmental point in the past. We could even understand the diachronic version in evolutionary terms, as the claim that the evolutionary emergence of the rational faculty was a necessary (and sufficient) prerequisite to the emergence of human moral judgment.

I want to stress that I am not taking it for granted that it is obvious what the terms “rational faculty” and “emotional faculty” mean. There is room for a great deal of discussion on this topic, but not in this paper. Here I am willing to employ these terms in a rough and ready way, since my intention is just to clarify the dialectic at a very broad level. I concede that it is possible that empirical science (including neuroscience) will cast into doubt the very idea of there being a “rational faculty” and an “emotional faculty.” If so, then all versions of Psychological Rationalism will be shown to be founded on an empirical misconception. If, on the other hand, the rational/emotional faculty dichotomy turns out to be broadly scientifically respectable (as I assume in what follows), then it’s an empirical issue—and in part a neuroscientific issue—whether a certain phenomenon (moral deliberation) causally involves one or the other, or both, of these faculties.

But we should be aware of how tricky it might be to undermine Psychological Rationalism. Merely observing an enormous amount of emotional activity when subjects engage in moral thinking is insufficient to sink any version of the theory, for I’d be surprised if any of the historical supporters of rationalism—even supporters of strong synchronic Psychological Rationalism—would object to the claim that of course moral thinking engages our passions. I doubt that Plato or Aquinas or Kant (another conspicuous moral rationalist) would have been in the least surprised to learn that neuroimaging reveals a great deal of emotional activity occurring in subjects when they are asked to contemplate hiring someone to tie up and rape their wife, or selling their young daughter to a pornographer (two examples

---

<sup>4</sup> What I am calling “Psychological Rationalism” Shaun Nichols (2002, 2004) calls “Empirical Rationalism.” I prefer my label since it emphasizes that this is the tradition of moral rationalism that one finds in the field of psychology. Were it not for the ugliness of the word, I think “Facultative Rationalism” would be a good label.

from the fMRI study done by Josh Greene and colleagues). A great deal of the moral realm concerns actions and persons who prompt anger, or indignation, or disgust, or sympathy; and no plausible version of moral rationalism denies this fact. The eighteenth-century moral rationalist Richard Price wrote: “Some impressions of pleasure and pain, satisfaction or disgust, generally attend our perceptions of virtue and vice. But these are merely their effects and concomitants, and not the perceptions themselves” ([1758] 1974: 44). The supporter of strong synchronic Psychological Rationalism can allow that activity of the rational faculty—though necessary and sufficient for moral judgment—nevertheless on many occasions (or as a matter of fact always) is accompanied by emotional excitement. In order to refute this theory it is not enough to observe that moral deliberation is reliably attended by emotional activity; one would have to show that moral deliberation is not also accompanied by rational activity. Neuroimaging could in principle yield this result, though as far as I know it hasn’t yet. (One major challenge to such a project would be first to operationalize the occurrence of “activity in the rational faculty.”)

Another conspicuous test procedure would be to investigate subjects who have various kinds of rational and/or emotional impairment, and see how this affects their capacity to engage in moral deliberation. Suppose, first, that we were to locate subjects perfectly able to make moral judgments but who suffer impaired rational faculties. We might be tempted to conclude on this basis that rational activity is not necessary for moral judgment, and thus that both strong and weak synchronic Psychological Rationalism stand refuted. But this strategy is undermined by the general reflection that a defective system may nevertheless continue to yield undamaged outputs of a certain sort—just as a faulty memory faculty may nevertheless recall certain events with great clarity, or a faulty clock may nevertheless reliably convey the accurate date. It is quite possible that moral judgment is the product of the rational faculty, but that this faculty can suffer certain forms of impairment while still merrily turning out well-formed moral judgments. Strictly speaking, what we would need to observe in order to refute the claim that rational activity is necessary for moral judgment is a subject who continues to make moral judgments despite having *no* rational activity. But it is unclear what criteria would need to be fulfilled before we were satisfied that we had such a subject. (Having *no* rational activity, I’m sure it will be agreed, is a pretty severe affliction.) Furthermore, such evidence would still leave diachronic versions of Psychological Rationalism viable, for even if moral judgment is possible without rational activity occurring there and then, nevertheless it is possible that activity of a rational faculty was necessary at some earlier point in time.

We might, on the other hand, locate a second kind of subject: one who has intact rational faculties but who appears unable to engage properly in moral thinking. Such cases would seem to show that rational activity is not (synchronically) sufficient for moral judgment—thus refuting strong synchronic Psychological Rationalism. Apparently we do have at least this kind of evidence. Shaun Nichols (2002, 2004) argues that psychopaths represent a class of persons whose rational faculties are intact, whose emotional faculties are impaired, and whose capacity to engage in moral deliberation is defective (see also Blair 1995; Blair et al. 1997). Assuming that this is an accurate description of the phenomenon of psychopathy, these subjects pose a serious challenge for strong synchronic Psychological Rationalism. The weak synchronic Psychological Rationalist, however, need not be troubled by the phenomenon of psychopathy, since it is consistent with this evidence that activity of the rational faculty remains necessary for moral judgment.

*Conceptual Rationalism* is the view that a reference to practical rationality will appear in any adequate explication of our moral concepts: that it is a conceptual truth that moral transgressions are transgressions of practical rationality. Though I should not like to exclude the possibility of neuroscience having an influence on what conclusions should be drawn about the content of concepts, it must be confessed that it is difficult to see how that contribution might transpire. Of course, the matter of how one *should* proceed to uncover the content of concepts is a very good question, about which philosophers argue. In so far as uncovering concepts means figuring out what we mean by the words we use to express those concepts, and figuring out what we mean by the words we use is a matter of figuring out how we *use* those words, and figuring out how we use words is an empirical matter, then uncovering the content of concepts is an empirical matter. But it is difficult to see what neuroscience in particular could contribute to this empirical process.

Nichols (2002, 2004) has performed an empirical survey of people's intuitions which, he argues, casts Conceptual Rationalism into doubt. His argument has a two-part structure. First he suggests that Conceptual Rationalists are committed to the following thesis:

SIMPLE MOTIVATION INTERNALISM:<sup>5</sup>

Anyone who judges that she is morally required to  $\phi$  will be motivated to comply.

Given this first step of arguing that Conceptual Rationalism implies Simple Motivation Internalism, a natural way of attacking Conceptual Rationalism would be to argue that Simple Motivation Internalism is false. But this strategy—which I will come to in a moment—is not actually Nichols's line of attack. Rather, his second step is to argue that ordinary people readily admit the existence of persons who represent counter-examples to Motivation Internalism, and thus the status of Motivation Internalism as a conceptual truth is doubtful. Nichols conducted an experiment where subjects were given a description of a psychopathic individual, John, who claims to know the difference between right and wrong while remaining utterly unmotivated to act accordingly. Then the subjects were asked whether John really made moral judgments. Most subjects maintained that he did. Thus Nichols concludes that it appears to be a "folk platitude that psychopaths understand that it is morally wrong to hurt others but don't care." It is important to note that Nichols doesn't think that real psychopaths actually do represent a counter-example to Motivation Internalism—as I mentioned before, it would appear that their capacity to engage in moral deliberation is fairly defective. Nichols's argument is simply that most people naively and perhaps erroneously believe that psychopaths represent a counter-example to Motivation Internalism, and thus Motivation Internalism cannot be a conceptual truth.

What are we to make of this argument? First I should remind you that it is not my concern to defend Conceptual Rationalism—for all I care it may be false. My intention is just to identify what bears on it and what doesn't. And since Nichols's evidence against Conceptual Rationalism is in no sense neuroscientific—but rather just concerns what ordinary people are likely to say about psychopaths—it's not strictly within my purview. Nevertheless, I want to make a few brief comments, since they do bear directly on what follows.

---

<sup>5</sup> Following Michael Smith (1994), Nichols actually calls this thesis "the Practicality Requirement," though I will defer to metaethical tradition and call it "Motivation Internalism," qualifying it as "simple" in order to contrast it with variants to be discussed shortly.

The moral rationalist that Nichols has most clearly in his sights is Michael Smith (1994). Smith certainly is a Conceptual Rationalist, and also endorses a form of Moral Internalism, and also thinks that the former supports the latter. However, we need to look carefully at the version of Moral Internalism that Smith endorses, for it isn't this simple variety. Smith doesn't think that moral judgment *guarantees* motivation; his version of Motivation Internalism is altogether more normative:

SMITH'S NORMATIVE MOTIVATION INTERNALISM:

Anyone who judges that she is morally required to  $\phi$  will be motivated to comply, or she is irrational.

Elsewhere Smith says that a person making a moral judgment will be motivated to comply "absent the distorting influences of weakness of will and other similar forms of practical unreason" (1994: 61), and these other forms of practical unreason are listed elsewhere in his book as "psychological compulsions, physical addictions, emotional disturbances, depression, spiritual tiredness, accidie, illness and the like" (1994: 154). This clarification suggests the following thesis, which is best interpreted as an explication of (rather than an alternative to) the former:

SMITH'S SUBSTANTIVE MOTIVATION INTERNALISM:

Anyone who judges that she is morally required to  $\phi$  will be motivated to comply, absent the distorting influences of weakness of will, psychological compulsions, physical addictions, emotional disturbances, depression, spiritual tiredness, accidie, illness, and the like.

The main problem with Nichols's empirical test concerning people's views about psychopaths is that it apparently targets *neither* of Smith's versions of Motivation Internalism. The subjects were not asked whether John the imaginary psychopath might be suffering from weakness of will or spiritual tiredness, nor whether he might be accused of irrationality for remaining unmoved by his moral judgments. All the test shows is that people readily countenance the falsity of *Simple* Motivation Internalism. But Smith never denied that.

A second thing that can be said in Smith's defense is that he has a rather distinctive view of what a conceptual truth is. Conceptual truths, for Smith, can be terribly unobvious to ordinary speakers. To have competence with a concept is to know *how* to use the word that stands for the concept—and that know-how may be very difficult to articulate even for the people who have it. By analogy, it might be very bad way of figuring out how a champion swimmer swims to ask him to describe his own swimming technique. Hence Smith is not terribly impressed with questionnaires revealing people's intuitive responses to set questions. Such questionnaires surely have *some* bearing on conceptual content, but they're a long way from settling the matter. If you want to know the content of a concept, then the best person to ask—Smith thinks—is an expert who has examined the patterns of usage of moral language as it is employed in real life.

So Smith is not forced to retreat from his Conceptual Rationalism in the face of Nichols's empirical evidence. Let me turn now to a different form of argument. I mentioned earlier the possibility of another kind of case that might be made against Conceptual Rationalism, this one also starting with the first step of showing that Conceptual Rationalism implies Simple Motivation Internalism, and then making the second step of showing that Simple Motivation Internalism is false. This argument is of interest to us because it has been claimed that

neuroscientific evidence does indeed reveal Simple Motivation Internalism to be false. Adina Roskies (2003) offers the evidence of patients suffering from localized injury to the ventromedial cortex, who appear to make normal moral judgments while remaining utterly disinclined to act accordingly. Damasio and colleagues (1990) have referred to this phenomenon as “acquired sociopathy.”

It must be emphasized that Roskies is not out to attack Conceptual Rationalism—her target is just Simple Motivation Internalism. However, if the link from Conceptual Rationalism is in place, then one could use her results to mount such an attack. In what follows I will argue that both steps of this argument fail: The link from Conceptual Rationalism to Simple Motivation Internalism can be severed (thus severing any *modus tollens* link running in the other direction), and, in any event, the empirical case against even Simple Motivation Internalism is flawed.

Let’s first look more carefully at this supposed implication from Conceptual Rationalism to Motivation Internalism. Nichols offers no discussion of this point, content to note that “the most prominent and influential versions of Conceptual Rationalism are tied to [Motivation Internalism], and I will simply assume in what follows that Conceptual Rationalism is committed to [Motivation Internalism]” (2004: 72 n.3). But this implication is not to be assumed without comment, for it is obvious that in order to move from Conceptual Rationalism as a premise to Motivation Internalism as a conclusion, an additional bridging premise is needed—one that links rational requirements to motivation. I’ll call the thesis “Rational Internalism.”

- Premise 1:*           CONCEPTUAL RATIONALISM:  
                          To judge that one is morally required to  $\phi$  is to judge that one is rationally required to  $\phi$ .
- Premise 2:*           SIMPLE RATIONAL INTERNALISM:  
                          Anyone who judges that she is rationally required to  $\phi$  will be motivated to comply.
- Therefore:*           SIMPLE MOTIVATION INTERNALISM:  
                          Anyone who judges that she is morally required to  $\phi$  will be motivated to comply.

Smith certainly endorses a version of Rational Internalism, which allows him to move from Conceptual Rationalism to Motivation Internalism. But, again, it isn’t this simple form that he endorses—as before, it is a normative version that he argues for: that anyone who judges that she is rationally required to  $\phi$  will be motivated to comply *or she is irrational*. He argues at some length for the view that what it is to judge that some action is rationally required of you is to believe that a fully rational version of yourself would desire that you do that thing. For someone—say, Fred—to judge that practical rationality is on the side of eating another slice of pizza is for Fred to judge that an idealized version of himself—that is, idealized in the respects of being granted full information and perfect powers of reflection—would advise the actual less-than-ideal Fred to have another slice of pizza. Suppose, then, that Fred does indeed believe this: that he would desire another slice of pizza if he were fully rational, but in fact doesn’t desire another slice of pizza. Is he irrational? Smith answers: “Most certainly.” For Fred has failed to have a desire that by his own lights it is rational for him to have.

Notice, though, that with a normative version of Rational Internalism the only version of Motivation Internalism that may be validly derived is also the normative variety, yielding this argument:

*Premise 1:* CONCEPTUAL RATIONALISM:

To judge that one is morally required to  $\phi$  is to judge that one is rationally required to  $\phi$ .

*Premise 2:* NORMATIVE RATIONAL INTERNALISM:

Anyone who judges that she is rationally required to  $\phi$  will be motivated to comply, or she is irrational.

*Therefore:* NORMATIVE MOTIVATION INTERNALISM:

Anyone who judges that she is morally required to  $\phi$  will be motivated to comply, or she is irrational.

But this kind of normative Motivation Internalism is one that Roskies admits the empirical evidence doesn't refute. In fact, she disparages Smith's brand of Motivation Internalism as "too weak to be revealing about the nature of moral judgment" since, she thinks, it leaves entirely open what is to count as irrationality (53). We've already seen that this charge is a bit unfair on Smith, for he does provide an inventory of phenomena that are supposed to jointly constitute *practical irrationality*. It's true that his list ends with a worrying "... and the like," but he doesn't leave matters entirely open. In any case, subjects suffering from acquired sociopathy are unlikely to represent counterexamples to Smith's Motivation Internalism, since it is doubtful that we will describe them as free from "emotional disturbance" or "illness." I strongly suspect Smith would be willing to add "brain damaged" to his list.

If we want Simple Motivation Internalism as our conclusion, then validity requires that we have a matching version of Simple Rational Internalism as the bridging premise. (This argument is presented above.) But the problem now is that such a version of Rational Internalism doesn't look like a good contender for an obvious truth at all; certainly there is nothing in Smith's work supporting such a standpoint. Moreover, it seems that many of the reasons that would lead one to doubt Simple Motivation Internalism will also lead one to doubt the truth of Simple Rational Internalism. If we're willing to acknowledge the existence of agents whose motivational structures are so impaired that we can imagine them sincerely saying "Yes, I know that I morally ought to  $\phi$ , but I just lack any motivation in favor of  $\phi$ -ing," then what is there to stand in the way of us also acknowledging motivational impairment that leaves an agent sincerely saying "Yes, I know that I rationally ought to  $\phi$ , but I just lack any motivation in favor of  $\phi$ -ing"?

In other words, Motivation Internalism and Rational Internalism look like they will stand or fall together. But this is good news for the Conceptual Rationalist. Simple Motivation Internalism may be false, but if the matching Rational Internalism is also false then the link to Conceptual Rationalism is severed. Or Normative Rational Internalism may be true, but if the only version of Motivation Internalism that may thus be derived is also the true kind, then again the Conceptual Rationalist has nothing to worry about. At the very least, once we acknowledge that Conceptual Rationalism alone doesn't imply Moral Internalism, but requires another substantive bridging premise, then we must admit that the falsity of Moral

Internalism is insufficient to sink Conceptual Rationalism, since this thesis can always be saved by dumping the bridging premise instead.

I noted that Roskies's target is not Conceptual Rationalism, but just Simple Motivation Internalism. But doubts may also be raised as to whether the empirical evidence she cites even has any impact on the well-entrenched metaethical dispute over Motivation Internalism. In order for those subjects suffering from acquired sociopathy to count as "walking counterexamples to this internalist thesis," as Roskies claims they are, two things need be true of them: They must make moral judgments and they must have no motivation to comply. The second requirement is something that can, to a reasonable extent, be empirically operationalized: Roskies cites the subjects' flat skin-conductance responses when faced with emotionally-charged or value-laden stimuli, supported by clinical histories indicating motivational impairments. The problem is that the criteria for having made a moral judgment are not similarly operationalized. The very notion of "a moral judgment" is sufficiently indeterminate that the conspicuous possibility is that there is a legitimate sense in which these characters are making moral judgments and an equally legitimate sense in which they're not. My suspicion is that it is precisely this indeterminacy that explains why the cottage industry in metaethics regarding Motivation Internalism has been so long locked at this tedious impasse.

If we treat moral judgment primarily as a kind of linguistic performance—as a speech act—then it is indeed reasonable to assume these patients capable of making moral judgments (though see Kennett, this volume). But, of course, when we treat moral judgment this way then we hardly need an appeal to modern neuroscience to refute the internalist thesis, for such moral judgments lacking motivation are no more exceptional than insincere apologies. Picture the situation of a person trying to curry favor with another by agreeing with all her evaluative assessments. Perhaps a young man decries the plight of farm animals in order to impress his new vegetarian sweetheart, despite the fact that really he couldn't care less. If moral judgments are considered just as speech acts, then our besotted young pretender has surely made one (albeit insincerely) while lacking any motivation to comply. The existence of such a phenomenon is unremarkable.

If, on the other hand, we prefer to treat moral judgment as more of a psychological event—as a kind of internal "mental assent" to an evaluative proposition—then serious doubt arises as to whether the subjects suffering from acquired sociopathy really are making moral judgments in this more robust sense.

This dilemma may be clarified by again comparing the case of apologizing. Apologies considered as speech acts can of course be insincere—they can be successfully performed by persons lacking any regret. But we might on occasion prefer to speak of apologies in a different sense: We may want to use the term to denote the type of mental event that occurs in persons' minds when they truly accept their apology, when they sincerely have regret and genuinely acknowledge responsibility "in their heart" (so to speak). Evidence that people can make apologies sans regret *in the former sense* clearly does not amount to evidence that people can make apologies sans regret *in the latter sense*. If there were a type of brain damage that left people incapable of feeling genuine regret, this wouldn't necessarily leave them lacking the very concept of an apology. Such people may know what apologies are, and may even learn when to say "Sorry" in appropriate circumstances in order to avoid social exclusion, but we may well decide that there's an important sense in which they are unable to apologize *sincerely*.

It is not clear what sense of “moral judgment” Roskies has in mind. She emphasizes the subjects’ linguistic skills—pointing out that they have “mastery of moral terms” and that “their language and declarative knowledge structures are intact” (60)—and this might be taken to suggest that she thinks of moral judgment as a linguistic performance—in which case Motivation Internalism surely stands refuted (but we hardly needed any fancy empirical evidence to demonstrate this). On the other hand, perhaps she is treating this linguistic competence merely as evidence of some kind of mentalistic moral judgment. But why should we accept that it does count as evidence? One might be tempted to say that linguistic competence is evidence of conceptual competence, and thus the subjects suffering from acquired sociopathy retain mastery of the moral concepts—when they say “Killing is morally wrong” they know what they are saying. And thus (one might be tempted to add) if such persons are able to apply the moral concepts to the appropriate kinds of item (e.g., not to pieces of furniture or days of the week), then despite their impairments they retain the capacity to make moral judgments in the psychological sense indicated.

But these temptations will appeal only to those who already harbor cognitivist leanings; the moral noncognitivist, by contrast, will remain unimpressed and untempted. The traditional noncognitivist—one version of whom is the emotivist—denies that moral judgment consists of applying concepts, and, indeed, to the extent that he claims that the predicates “...is morally good,” “...is morally prohibited,” etc. are only grammatical predicates but are not logical predicates, then he may even deny that there exist such concepts as *moral wrongness*, *moral requirement*, etc. (in the same way as one might deny that *yum!* or *boo!* count properly as concepts). Suppose, for example, that there are well-entrenched linguistic conventions according to which when one makes a public moral judgment one thereby expresses subscription to the relevant normative framework (see Gibbard 1990). If this is correct, then to make a moral judgment in the psychological sense will involve a sincere subscription to the norms indicated by the public judgment. Since such “subscription” is (*ex hypothesi*) a motivation-implicating state, it follows that moral judgment *in this sense* implies motivation. Such a noncognitivist, accordingly, treats these subjects suffering from acquired sociopathy not as counterexamples to Motivation Internalism, but as counterexamples to the proposition that mastery of the moral language suffices for the occurrence of moral judgment.

Roskies, it would seem, has made herself immune to such complaints by putting to one side the possibility that moral judgments may function to do anything other than express beliefs. She writes: “For the purposes of this paper, I will assume moral cognitivism to be true” (53), and admits that “[t]he arguments presented here do not suffice to refute internalism in a non-cognitivist framework” (64). This is curious given that the biggest fans of Motivation Internalism in metaethics have traditionally been the noncognitivists, and indeed much of the philosophical interest in the thesis lies in the supposition that its resolution promises to shed light on the cognitivist/noncognitivist dispute (though whether this is a reasonable expectation is moot—see Joyce 2002, forthcoming *a*). Most moral philosophers who embrace pure cognitivism—the view that the linguistic function of a moral judgment is exhausted by its belief-expressing quality—see Motivation Internalism as an unlikely and unnecessary thesis. This assessment arises not from any neuroscientific research, but generally from reflection on the Humean psychological thesis that beliefs and desires are but contingently linked entities.<sup>6</sup>

---

<sup>6</sup> Most moral cognitivists who embrace Simple Motivation Internalism (such as John McDowell) adopt some fairly unorthodox views about beliefs and desires in order to square things. (See Kennett, this volume.) In fact,

In other words, once the conditionality of Roskies's argument is highlighted, much of the metaethical interest in it evaporates.

Over the past few paragraphs I have been casting doubt on the supposition that neuroscientific evidence sheds light on the entrenched metaethical debate over Simple Motivation Internalism. I have argued neither that this internalist thesis is true nor that it is false. My point is rather that the notion of a moral judgment is sufficiently pliable to allow of reasonable precisifications according to which internalism is pretty obviously false, and equally reasonable precisifications according to which it may be true. The latter depends on whether the "sincere acceptance" of a moral judgment implicates motivational structures, which in turn depends on whether there exist linguistic conventions according to which public moral judgments function to express (*inter alia*, perhaps) conative attitudes. To the extent that this is an empirical matter, it is a job for socio-linguistics; I see no obvious place for a significant contribution from neuroscience. This is a noteworthy subconclusion in its own right, but my broader aim has been to undermine the claim that neuroscientific evidence establishes a premise (the falsity of Simple Motivation Internalism) that might be used in an attack on Conceptual Rationalism.

Let me turn finally to *Justificatory Rationalism*. According to this view, moral transgressions are rational transgressions; moral villains are irrational. This is distinct from the claim made by the Conceptual Rationalist since it doesn't assert a conceptual connection. This is important, for it makes it all the clearer that collating people's intuitions on the issue (in the manner of Nichols's above experiment) has little bearing on the matter.

In order to gain a rough idea of how this kind of moral rationalist thinks the justificatory process may proceed, let us briefly consider the views of Peter Singer. Singer argues that natural selection has granted humans an innate tendency to look favorably upon actions that benefit one's family, and a tendency to dislike actions that harm them. However, Singer goes on to argue, we have also been granted by natural selection a rational faculty. Rationality allows a person to realize:

I am just one being among others, with interests and desires like others. I have a personal perspective on the world, from which my interests are at the front and center of the stage, the interests of my family and friends are close behind, and the interests of strangers are pushed to the back and sides. But reason enables me to see that others have similarly subjective perspectives, and that from "the point of view of the universe" my perspective is no more privileged than theirs. (Singer 1995: 229)

---

however, if one is treating moral judgments as speech acts it is perfectly possible that they might be necessarily linked to motivation-implicating states without functioning to express such states. And one need not deny Humean views on beliefs and desires in order to acknowledge this. Consider yet again the act of apologizing. The criteria for an apology to have occurred involve the need for both parties to be versed in a range of relevant linguistic conventions; for example, the addressee must hear and understand the words uttered, and the speaker must take it that this is the case. The satisfaction of these criteria will require both speaker and addressee to have certain *beliefs*; for example, the speaker must believe that his addressee hears and understands. This connection is a necessary (and *a priori*) one: It is not possible that any person could succeed in apologizing to another person without having such a belief. Yet we would hardly say that the act of apologizing—the utterance of "I'm sorry" in the appropriate circumstances—*functions to express the belief that one's audience hears and understands*. This suffices to show that the occurrence of a type of speech act may entail that the speaker has a certain kind of mental state, though the speech act doesn't function to express that state. Thus moral cognitivism and Simple Motivation Internalism may be compatible without denying Humean psychology. See Joyce 2002.

Reason, Singer seems to be saying here, demands that one recognize the welfare of others as being as objectively valuable as one's own welfare. It wouldn't follow from this that we all need to be totally impartial in our actions—showing no favor to friends and family—for that might lead to a very unhappy state of affairs; but it is at least supposed to show that, for instance, it is unacceptable to go up to an innocent stranger and punch him for my amusement. Given that I would demand not to be punched, I am rationally required to acknowledge that there is an equal demand that the stranger not be punched. Regardless of how plausible we find this view (I certainly have few sympathies with such thoughts<sup>7</sup>), I think it fair to say that this is the core intuition motivating many moral rationalists: Michael Smith, Christine Korsgaard, Thomas Nagel, Alan Gewirth, and, indeed, Kant. I would go so far as to call it *the* dominant thread of Western moral rationalism. Simon Blackburn has described the rationalist's goal of finding a “knock-down argument that people who are nasty and unpleasant ... are above all *unreasonable*”—that such villains “aren't just selfish or thoughtless or malignant or imprudent, but are reasoning badly”—as the “holy grail of moral philosophy” (Blackburn 1984: 222).

The main point I wish to stress is the distinction between Justificatory Rationalism and Psychological Rationalism, and the simplest way of achieving this is to point out that Justificatory Rationalism primarily concerns *action* not judgment, whereas Psychological Rationalism concerns the sources of moral judgment.<sup>8</sup> Let us say that a certain person is morally required to give to a particular charity, and suppose that she does so. The Justificatory Rationalist's purview extends only to the claim that she would have been irrational had she refrained from giving to the charity. As to the question of what may have motivated her action, the Justificatory Rationalist is silent. Perhaps she didn't do it for a *moral* consideration at all. Or even if she did do it because she thought it morally obligatory, the Justificatory Rationalist is silent about the source of this moral judgment. Perhaps it springs from her rational faculty, or perhaps it flows from seething emotional activity—and emotional activity alone—or perhaps, as Haidt argues (this volume), her judgment is a post-hoc construction to some knee-jerk emotional response. The Justificatory Rationalist is not saying that moral judgments always, or even typically, (or even, in principle, *ever*) causally flow from the proximate activity of a rational faculty, but rather that the principles of rationality favor a certain degree of impartiality in our dealings with each other, from which it follows that a person's rational faculty would, if properly exercised and unimpeded, recognize this fact. But perhaps the properly exercised and unimpeded rational faculty is a rare thing.

It is possible that Singer does see a causal link between the rational faculty and moral judgment in evolutionary terms. Perhaps if our ancestors had never evolved the sophisticated rational abilities that humans presently do enjoy then we'd never have gotten beyond liking actions that help ourselves and our kin, and disliking actions that harm them, in which case perhaps we'd never have started making *moral* judgments at all. And in this light Singer

---

<sup>7</sup> See my 2001: chapters 4 and 5.

<sup>8</sup> There is, of course, a natural bridge from practice to judgment: If punching someone is rationally unjustified then it surely follows that the decision/judgment to refrain from punching that person could be justified by an appeal to the principles of practical rationality. Nevertheless, this observation doesn't imply that Justificatory Rationalism has anything to say about *moral* judgment, for one can justify an action by appealing to practical rationality without making a moral judgment: Given my present desires, it may be practically rational for me to pause in writing this footnote in order to make a cup of tea—and I could surely justify this decision by an appeal to the principles of practical rationality—but no *moral* judgment would figure in my deliberations.

might be interpreted as a kind of Psychological Rationalist (as, indeed, Nichols interprets him [2004: 68]), but it is vital to note that to the extent that this is a reasonable interpretation it is a historical diachronic version of Psychological Rationalism that Singer is pressing. He is not arguing that all (or even many) moral judgments are caused by the operations of the rational faculty of the person making the judgment. Even for those persons who do not properly exercise their rational faculty, who act selfishly and show extreme partiality towards their kith and kin—bearing in mind that this may be nearly all of us, nearly all the time—it nevertheless may be that a degree of impartial benevolence in their actions *is* rationally required, and thus such people are *being* irrational. To the extent that Singer endorses this claim (his views on this matter in fact have a complexity to which I am unable to do justice on this occasion), it is a Justificatory Rationalism that he advocates.

And it is this consideration that also makes it obvious that the truth (or otherwise) of Justificatory Rationalism will not be affected by neuroscientific research concerning what is going on in people's brains when they make moral judgments, for the theory is compatible with just about any discovery concerning the springs of moral judgment and action. All that is required of human psychology in order for Justificatory Rationalism to be reasonable is that we at least fulfill the minimal requirements for being rational agents, for I take it that few would support the view that creatures constitutionally incapable of complying with rational considerations *as such* can still be subject to rational requirements. Lions, for instance, cannot be accused of irrationality (or immorality) for failing to adequately take into account the welfare of gazelles. But this is a very modest constraint, and one which we can be pretty confident will not be affected by neuroscientific advances. Even if neuroscience were to scotch the idea that there is anything like a rational *faculty* in the brain, we could still claim to have the skills sufficient for being rational beings. Even if empirical data were to show that irrationality pervades our decisions, this would (ironically) presuppose that humans fulfill the prerequisites for being rational in some more broad sense of the term, for only beings with the capacity for rationality can be properly criticized as acting and thinking irrationally (which is why we don't accuse lions of being irrational).

Certain generic descriptions of moral rationalism may encourage one to overlook the fundamental difference between Psychological and Justificatory Rationalism. Asserting that moral judgments “derive from” rationality is ambiguous (as too is the phrase “have their source in”). It can be read etiologically, as concerning the actual mechanisms that produce moral judgments—as Haidt does when he describes the rationalist as holding that “moral knowledge and moral judgment are reached primarily by a process of reasoning and reflection” (2001: 814). Or it can be read normatively, as concerning the principles that underwrite and justify the contents of moral judgments, regardless of the causal source of those judgments. Both kinds of rationalist can legitimately claim to be seeking “the foundations” of morality, but they are in fact engaged in very different pursuits. The empirical evidence may refute Psychological Rationalism—and neuroscience may contribute to its downfall—without this in any way compromising the moral rationalist's project of finding a rational foundation for ethics.

### **3. A more optimistic conclusion?**

I have here undertaken the modest task of sorting out a few potential cross-disciplinary confusions, some of which are purely terminological. Given that such confusions can lead to misunderstanding and wasted academic effort, this seems a worthy task to perform at a time when the empirical sciences have begun to enrich moral philosophy in ways that were not anticipated a generation ago, but which we can now expect will burgeon in coming years. Because the tone of this chapter has largely been negative about the contributions neuroscience may offer metaethics, I want to stress that there is no cause for general pessimism regarding the possibility that such contributions may be forthcoming in ways that have not been discussed. I am, in fact, confident that empirical data—including that of a neuroscientific nature—will impact on a number of metaethical issues. My guess is that the greatest contribution will be to moral epistemology, though perhaps in ways that will be found unsettling (see my 2006a: chapter 6, 2006b). Walter Sinnott-Armstrong (2005), for example, argues that empirical psychology reveals many moral judgments to have attributes that, by ordinary epistemic standards, render them in need of independent confirmation. Whatever privileged status we might have otherwise accorded a held belief (based on a principle of epistemic conservatism, say) is undercut if we discover that belief to be partial, controversial, clouded by emotion, subject to illusion, or explicable by unreliable or disreputable sources. Certainly neuroscience and social psychology have revealed moral deliberation frequently to be clouded by emotion (even in cases when we wouldn't ordinarily think so)—so this is an instance of empirical data having a direct metaethical pay-off. Neuroscience may also have a role to play in establishing that the human moral faculty is innate (in some sense of the word). It is not unlikely that a picture will emerge of an innate moral faculty which will at no point presuppose moral thinking to have evolved in order to detect a realm of moral facts, but rather proposes that such thinking enhanced our ancestors' reproductive fitness via soothing and reinforcing prosocial relations (see my 2006a, 2006c). Since such a hypothesis concerning the genealogy of human moral judgment would nowhere presuppose that any such judgment is true, one might very well claim that this would amount to an empirical confirmation that moral judgments satisfy Sinnott-Armstrong's final criterion: of deriving from "an unreliable source."<sup>9</sup>

I said that there's no need to be pessimistic concerning the possibility of the empirical sciences contributing to metaethics, though one may very well find this particular kind of positive contribution pessimistic (in another sense): if one dislikes the direction in which the argument leads us. The general worry is that empirical discoveries about the genealogy of moral judgments may undermine their epistemic status and ultimately detract from their authoritative role in our practical deliberations. This is a possibility to be taken very seriously and explored carefully.<sup>10</sup>

---

<sup>9</sup> It is worth comparing this for clarity with a different psychological phenomenon for which an evolutionary hypothesis seems plausible: humans' simple arithmetic skills. Would the fact that we have such a genealogical explanation of our simple mathematical judgments serve to demonstrate that they are the product of an unreliable source? Surely not, for false mathematical judgments just aren't going to be very useful: Being chased by three lions, you observe two quit the chase and conclude that it's now safe to slow down. The *truth* of "1+1=2" is a background assumption to any reasonable hypothesis of how this belief might have enhanced reproductive fitness.

<sup>10</sup> Thanks to Adina Roskies, Peter Singer, Walter Sinnott-Armstrong, and Michael Smith for comments.

## BIBLIOGRAPHY

Ayer, A.J. [1936] 1971. *Language, Truth and Logic*. Penguin Books.

Blackburn, S. 1984. *Spreading the Word*. Oxford: Clarendon Press.

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

Blair, R.J.R., Jones, L., Clark, F. & Smith, M. 1997. "The psychopathic individual: A lack of responsiveness to distress cues?" *Psychophysiology* 34: 192-198.

Copp, D. 2001. "Realist-Expressivism: A neglected option for moral realism." *Social Philosophy and Policy* 18: 1-43.

Damasio, A.R., Tranel, D. & Damasio, H. 1990. "Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli." *Behavioral Brain Research* 41: 81-94.

Fessler, D.M.T., Arguello, A.P., Mekdara, J.M. & Macias, R. 2003. "Disgust sensitivity and meat consumption: A test of an emotivist account of moral vegetarianism." *Appetite* 41: 31-41.

Gibbard, A. 1990. *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M. & Cohen, J.D. 2001. "An fMRI investigation of emotional engagement in moral judgment." *Science* 293: 2105-2108.

Greene, J.D. & Haidt, J. 2002. "How (and where) does moral judgment work?" *Trends in Cognitive Sciences* 6: 517-523.

Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M. & Cohen, J.D. 2004. "The neural bases of cognitive conflict and control in moral judgment." *Neuron* 44: 389-400.

Haidt, J. 2001. "The emotional dog and its rational tail: A social intuitionist approach to moral judgment." *Psychological Review* 108: 814-834.

Haidt, J. 2003. "The moral emotions." In R.J. Davidson, K.R. Scherer, & H.H. Goldsmith (eds.), *Handbook of Affective Sciences* (Oxford: Oxford University Press): 852-870.

Hume, D. [1740] 1978. *A Treatise of Human Nature*. L.A. Selby-Bigge (ed.), Oxford: Clarendon Press.

Hume, D. [1751] 1983. *An Enquiry Concerning the Principles of Morals*. Cambridge, MA: Hackett Publishing Company.

- Joyce, R. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.
- Joyce, R. 2002. "Expressivism and motivation internalism." *Analysis* 62: 336-344.
- Joyce, R. 2006a. *The Evolution of Morality*. MIT Press.
- Joyce, R. 2006b. "Metaethics and the empirical sciences." In J. Kennett and P. Gerrans (eds.), *Philosophical Explorations* 9: 133-148.
- Joyce, R. 2006c. "Is human morality innate?" In P. Carruthers, S. Lawrence & S. Stich (eds.), *The Innate Mind: Culture and Cognition*. Oxford University Press.
- Joyce, R. Forthcoming. "Expressivism, motivation internalism, and Hume." In C. Pigden (ed.), *Reason, Motivation, and Virtue* (Palgrave, MacMillan).
- Moll, J., de Oliveira-Souza, R., Eslinger, P.J., Bramati, I.E., Mourão-Miranda, J., Andreiuolo, P.A. & Pessoa, L. 2002. "The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic moral emotions." *Journal of Neuroscience* 22: 2730-2736.
- Nichols, S. 2002. "Is it irrational to be amoral? How psychopaths threaten moral rationalism." *The Monist* 85: 285-304.
- Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. NY: Oxford University Press.
- Price, R. [1758] 1974. *Review of the Principal Questions in Morals*. D.D. Raphael (ed.). Oxford: Clarendon Press.
- Roskies, A.L. 2003. "Are ethical judgments intrinsically motivational? Lessons from acquired sociopathy." *Philosophical Psychology* 16: 51-66.
- Singer, P. 1995. *How Are We to Live?* Amherst, NY: Prometheus Books.
- Sinnott-Armstrong, W. 2005. "Moral intuitionism meets empirical psychology." In T. Horgan & M. Timmons (eds.), *Metaethics After Moore* (NY: Oxford University Press).
- Smith, M. 1994. *The Moral Problem*. Oxford: Oxford University Press.