

Metaethics and the empirical sciences

Richard Joyce

[This is the penultimate draft of the article that appeared in *Philosophical Explorations* 9 (2006)
(special issue: *Empirical Research and the Nature of Moral Judgment*): 133-148.]

What contribution can the empirical sciences make to metaethics? This paper outlines an argument to a particular metaethical conclusion—that moral judgments are epistemically unjustified—that depends in large part on a posteriori premises.

Introduction

This special issue testifies to the fact that the empirical sciences have recently produced a great deal of data that is of interest to the moral philosopher, emerging from a wide range of disciplines. But does any of it actually contribute to the resolution of any of the moral philosopher's perennial questions? It is popularly thought that the answer to this must be negative on the grounds that one cannot derive an 'ought' from an 'is'. (This objection is frequently erroneously confused with what G.E. Moore in 1903 called 'the naturalistic fallacy' (see Bruening 1971).) But there are many questions in moral philosophy for which the answer will not amount to an 'ought' statement. Let us say (more for the sake of didacticism than accuracy) that to wonder what we morally ought to do in a given situation, or what kind of people we morally ought to be, is to engage in *ethical inquiry*. By contrast, to wonder about what it is we are doing when we engage in ethical inquiry—to wonder what it is we do when we wonder what it is we morally ought to do—is to engage in *metaethical inquiry*. More specifically, metaethics has been concerned with three interrelated projects: *semantic* questions about the nature of ethical discourse, *ontological* questions about the existence of, and nature of, moral facts, and *epistemological* questions concerning whether (and, if so, in virtue of what) moral claims may be justified. Since metaethicists are not in the business of issuing moral 'ought' claims (being more interested in the meaning of the word 'ought') the injunction against deriving 'ought' from 'is' has never applied to them.¹

In this paper I will outline one argumentative strand to the metaethical conclusion that moral judgments are epistemically unjustified. The argument is of interest here because its point of departure is a premise that can be established only empirically: that moral judgments have a certain kind of genealogy. This genealogy concerns the proximal etiology of moral judgment (e.g., the neuroscience of moral deliberation), the extreme distal etiology of the phenomenon (the evolution of the moral sense in the hominid lineage), and many points between (e.g., the ontogenetic emergence of the moral sense). The distinctive feature of this genealogy, I want to point out, is that it nowhere presupposes that the beliefs in question are true. My concerns in this paper do not require that I elaborate the details of this genealogy; rather, I simply want to address the question: 'When does knowledge of the genealogy of a belief undermine that belief?' Some philosophers are inclined to answer 'Never', having been taught that to assess the truth value of a belief on the grounds of its origins is to commit the 'genetic fallacy'. It is evident, however, that

the truth or falsity of a belief can *sometimes* be determined by the belief's origins. Consider the belief that no beliefs are innate. To discover that this is an innate belief would be to falsify it. More to the point is the observation that showing a belief to be *false* is not the only way to undermine it. Reflection on the origins of a belief may reveal the belief to be epistemically *unjustified* while leaving the question of its truth or falsity open

In the course of outlining how an empirically-supported genealogy may lead to this metaethical conclusion, I will pause to note two other metaethical premises that require substantive empirical input in order for the argument to proceed. The first concerns the extent to which the debate between the cognitivist and the noncognitivist is an empirical matter; the second concerns the contribution that empirical data can make towards establishing (or undermining) moral naturalism. As I say, space does not permit me to attempt the resolution of these matters; my goal is the less ambitious one of delineating the dialectic while attending to the places where philosophers must cease to proceed *au fauteuil*, and must enlist the help of their empirically-inclined colleagues.

Genealogical debunking

Indulge in a slightly silly thought experiment, and pretend there were such things as belief pills, such that taking one would inevitably lead to the forming of a certain particular belief (while at the same time invoking amnesia about the taking of the pill and about the existence of such pills in general). Suppose that there were a pill that makes you believe that Napoleon won Waterloo, and another one that makes you believe that he lost.² Suppose also that there were an antidote that can be taken for either pill. Now imagine that you are proceeding through life happily believing that Napoleon lost Waterloo (as, indeed, you are), and then you discover that at some point in your past someone slipped you a 'Napoleon lost Waterloo' belief pill. It is not a matter of your learning of the existence such pills and having no way of knowing whether you have ever taken one; rather, we're imagining that you somehow discover beyond any shred of doubt that your belief is the product of such a pill. Should this undermine your faith in your belief that Napoleon lost Waterloo? Of course it should. It doesn't show that the belief is *false*, but the knowledge that your belief is the product of a belief pill renders the belief unjustified (or perhaps shows that it was never justified in the first place, depending on one's epistemological tastes³), demanding that unless you can find some concrete evidence either in favor or against the belief you should cease to believe this thing—that is, you should take the antidote.

The intention of this make-believe scenario is to prime us for an analogical epistemological conclusion regarding moral judgment. Instead of Napoleon beliefs suppose it's our moral beliefs, and instead of belief pills suppose it is a complete and empirically-confirmed genealogy of these beliefs. If the analogy is reasonable, therefore, it would appear that once we become aware of this genealogy of morals we should (epistemically) do whatever is analogous to taking the antidote pill: cultivate agnosticism regarding all positive beliefs involving moral concepts until we find some solid evidence either for or against them. Note how radical this conclusion is. It is not a matter of allowing oneself to have an open mind about, say, the wrongness of abortion or the

rightness of canceling Third World debt; rather, it is a matter of maintaining an open mind about whether there exists *anything* that is morally right and wrong (as John Mackie argued (1977)).

But *is* the analogy fair? It may be objected that in the case of the belief pills the story has been carefully stipulated such that forming a belief as the result of taking a pill is entirely independent of whether or not the state of affairs necessary to render the belief true obtains in the world. But perhaps things stand differently for the genealogy of morals; perhaps the processes involved are likely to yield true beliefs. After all, in principle we can tell a genealogical story for *any* belief, but this fact obviously does not render every belief unjustified. Even if the genealogy in question were one that shows the beliefs to be innate (in some sense of the word), this would not necessarily epistemically debunk the belief. This is worth emphasizing for it is not unreasonable to suppose that the availability of an *evolutionary* genealogy is likelier to have a debunking effect than more proximal genealogical explanations, since we know that natural selection is a process for which practical success rather than accuracy is the summum bonum. It can be countered, however, that very often accuracy is the route to practical success. Consider, for example, our mathematical beliefs. There is some evidence that the distinct genealogy of these beliefs can be pushed right back into evolutionary history: that natural selection has provided humans with an inbuilt faculty for simple arithmetic (Butterworth 1999). Would the fact that we have such a genealogical explanation of a simple mathematical belief like '1 + 1 = 2' serve to demonstrate that we are unjustified in holding it? Surely not, for we have no grasp of how this belief might have enhanced reproductive fitness independent of assuming its truth. False mathematical beliefs just aren't going to be very useful. Suppose you are being chased by three lions, you observe two quit the chase, and you conclude that it is 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 come to be innate.⁴

As with arithmetic beliefs, there is a body of evidence that the distinct genealogy of moral beliefs (or perhaps just moral concepts) can be constructed on an evolutionary timescale: that the human moral sense is a discrete innate faculty owing its existence and nature to the social life of our ancestors hundreds of thousands of years ago (Darwin [1879] 2004; Alexander 1987; Katz 2001; Haidt and Joseph 2004; Joyce 2006). Can we make sense of its having been useful for our ancestors to form beliefs concerning *rightness* and *wrongness* independently of the existence of rightness and wrongness? Here I think the answer is a resounding 'Quite possibly'. Of course, to answer this properly we would have to examine the particular evolutionary hypotheses in detail, which lies beyond the scope of this paper. Suffice it to say that an acquaintance with the contemporary literature on the evolution of the human moral sense will reveal no background assumption that any actual moral rightness or wrongness existed in the ancestral environment. Whether we assume that the concepts *right* and *wrong* succeed in denoting properties in the world, or whether we think that they suffer from a referential failure that puts them on a par with the concepts *witch* and *ghost*, the plausibility of the hypotheses concerning how moral judgment evolved remains unaffected. Not so for the mathematical case: Were someone foolish enough to doubt that $1 + 1 = 2$, the plausibility of the evolutionary story concerning how having this belief enhanced our ancestors' fitness would evaporate. In other words, moral judgment might differ from arithmetic beliefs in being an instance where accuracy is *not* the route to practical success.

The same contrast should be drawn concerning the faculties of scientific inquiry that are in some sense products of biological natural selection, and which we deploy in conceiving and testing our evolutionary hypothesis. This observation deflects a concern raised by Peter Railton (2000) against the possibility of using evolutionary theory to undermine ethics. Railton claims that any argument moving from the empirical premise that human morality is the product of evolution to the conclusion that morality is thereby in some sense debunked ‘hammers itself into the same ground into which it had previously pounded morality’ (2000: 57), for the reason that the very faculties we employed in order to establish the empirical premise are themselves the product of natural selection. But I suggest that—as in the arithmetic example—we have no grasp of how any innate human faculties pertaining to ‘scientific inquiry’ might have been selected for independently of their producing judgments that at least have some positive connection to the truth. Thus the ‘evolutionary debunking of morality’ does not in this manner debunk itself.

That the evolutionary genealogy of morals may contrast with other cases (such as arithmetical and scientific beliefs)—in that it does not presuppose the truth of the beliefs—is a crucial observation. But it doesn’t *suffice* for establishing that for morality we have a debunking genealogy, for the possibility remains that an identity or a supervenience relation may hold between the items denoted in the genealogy and the moral properties represented in the belief’s content, in such a way that the genealogy renders the belief true after all. (I am making the terminological assumption that an identity or a supervenience connection between two things does not amount to a presupposition relation holding between the two things.) This will be explained further in a later section, where I will use a well-known argument by Gilbert Harman to frame the discussion. First, however, something needs to be said to address a natural objection to this whole line of argument: namely, that moral judgments do not involve *beliefs* at all and thus it is a mistake even to raise the question of their epistemic justification.

Noncognitivism

The noncognitivist argues that public moral judgments perform some function other than assertion. When one makes the claim ‘Lying to Sally was morally wrong’, for example, one is expressing some conative attitude towards this action—such as disapproval—or is commanding others to have this attitude, or is expressing subscription to a normative framework that condemns this act of lying, or performing some other non-assertoric function. The *pure* noncognitivist argues that moral language performs *only* this non-assertoric role; there is also room for a mixed theory allowing that moral language performs more than one function, one of which is the expression of beliefs (see Copp 2001; Joyce forthcoming *b*). Let us here focus on the kind of pure noncognitivism known as ‘emotivism’ (or ‘expressivism’), according to which the linguistic function of moral language is exhausted by the expression of some (specifiable) conative state, such as an emotion.

According to the traditional emotivist, there is no such thing as a moral belief. There are moral judgments, but these judgments do not express beliefs, but rather express some conative mental state. If this is correct, then the genealogical debunking strategy outlined in the previous section does not apply to morality, because the notion of *epistemic justification* at the heart of the

argument applies principally, and perhaps solely, to beliefs. After all, the initial intuition pump concerned *belief* pills, not *approval* pills. If moral judgments function to express emotion, it is not clear what framework we should use to assess their justification. Perhaps they should be deemed justified or unjustified according to some instrumentalist principle (see Campbell 1996), or perhaps they should not be assessed in this manner at all (as, say, headaches are not considered justified or unjustified).

Some modern emotivists argue that they can accommodate the existence of moral beliefs, moral assertions, and moral truth, while still retaining a distinctively emotivist framework (Blackburn 1984, 1993). If the modern emotivist is to continue the tradition of distinguishing his theory from opponents' by reference to belief, assertion, and truth, then distinctions must be drawn between minimal and robust versions of these phenomena. The emotivist might allow, for example, that moral judgments express beliefs according to some *minimal* standard of belief (thus avoiding the embarrassing view that there is no such thing as a moral belief), while maintaining that the beliefs in question are not 'full-bloodedly representational'.⁵ On the assumption that even these minimal beliefs are subject to assessment by whatever our general epistemological framework turns out to be, the pursuit of this emotivist strategy will leave moral judgments vulnerable to the genealogical debunking strategy.

Since the traditional ('belief-denying') kind of pure emotivism promises to evade the debunking argument, we need to ask how we might determine whether it is an acceptable theory. There are various ways in which one might be tempted to think that empirical research can decide the matter. In what follows I will mention but reject two such ways, and then identify a third body of empirical data that I believe does bear directly on the matter.

First, neuroscientists (Greene et al. 2001; Moll et al. 2003) and social psychologists (Haidt 2001) have provided evidence that engagement in moral deliberation centrally involves emotional activity, and this, one might be tempted to think, supports emotivism. But this temptation reveals a misunderstanding of the relevant 'expression' relation involved in the emotivist thesis that moral judgments *express* emotion: It is not a *causal* relation, but rather a matter of linguistic convention. By analogy, if a person asserts that the cat is on the mat, then he (by convention) has expressed the belief that the cat is on the mat; yet his assertion might be a lie, in which case he does not believe that the cat is on the mat, revealing that the expression of this belief cannot be a matter of the belief having *caused* the utterance. Thus, the most that psychological 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 moral judgments *express* those emotions. It is entirely possible that moral judgments are typically (or even always) caused by emotional activity but nevertheless function linguistically as assertions.⁶

Second, one might be tempted to think that evidence in support of certain evolutionary hypotheses about the emergence of the human moral sense lends credence to noncognitivism. Philip Kitcher (1998, 2005), for example, argues that if evidence indicates that moral judgments functioned evolutionarily to coordinate human social behavior, and if they achieved this by an 'amplification of our psychological altruistic dispositions' (2005: 178), then noncognitivism looks promising. His argument seems to be that if moral judgments function to augment some aspect of our emotional lives then 'the surface forms of moral judgments deceive us, [in that] we aren't really uttering straightforward declarative sentences but expressing emotive reactions'

(2005: 175). But it is important that we not conflate the evolutionary function of moral judgments with their linguistic function, as Kitcher appears to be doing. Perhaps Kitcher is correct that the evolutionary function of moral judgment is to generate or strengthen some kind of prosocial emotion; the question we must ask is ‘How does it accomplish this?’ Perhaps, for example, moral judgments encourage speakers (and their audience) to project their emotions onto the world in such a way as to think that there exist authority-independent practical demands, and conceiving of the world in these terms serves to ‘amplify altruistic dispositions’ in a fitness-advancing manner. Such an account recognizes that conative activity is crucial to moral judgment (by both prompting the projectivist activity and in turn being affected by it), but, importantly, it also implies an account of the phenomenology of those doing the projecting: It seems to them as if the world contains properties. Since we can assume that the language with which they discuss the matter will reflect their experience, then when they say things like ‘Lying to Sally was morally wrong’ it seems safe to assume, absent any reason to think otherwise, that they are expressing the belief that the act instantiates the property of wrongness. 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.

The kind of empirical evidence that does in fact bear on the debate between the cognitivist and the noncognitivist is nothing so exotic as neuroscience or evolutionary biology. Their debate concerns what kind of mental state(s) moral judgments express, where this ‘expression’ relation, as we have seen, is a matter of linguistic convention. Thus, what needs to be examined in detail are the conventions surrounding moral discourse in our natural language. Quiet and careful introspection is not a reliable source of knowledge on such matters; we really need to observe language being used across a wide range of everyday settings. (We might think of this empirical inquiry as ‘socio-linguistics’.) Note that a collective tendency to treat moral utterances as assertions is not merely good evidence that they are assertions, but is *constitutive* of their being so. The idea that a linguistic population might be unanimously inclined to treat a portion of their language as assertoric *but be mistaken about this* reveals confusion of what determines such facts.

Since my goal in this paper is to outline (rather than advocate) a metaethical argument, it is not my intention to take sides on these matters. In what follows I will simply assume that the traditional kind of pure emotivism (and noncognitivism) fails, and that therefore the normal standards of epistemic justification (whatever they are) apply to moral judgments. Let us, then, return to the question of genealogical debunking.

Harman’s challenge

The idea that moral judgments might be epistemically undermined on the grounds that they can be explained entirely without invoking their truth—i.e., without invoking any moral facts which they represent—has been discussed at some length by Harman (1977, 1986). Harman focuses not merely on whether moral judgments may be explained without invoking their truth, but on whether one need posit moral facts in order to explain *anything*. He is widely reported as having answered in the negative, but in fact his position is subtler. Rather, Harman’s conclusion is conditional: that *if* there is no reductive account available explaining how moral facts relate to

naturalistic facts, then moral claims cannot be tested, moral theories cannot be confirmed or disconfirmed, and we have no evidence for the existence of moral facts. But at no point does Harman assert that the antecedent of this conditional holds. In fact, he has himself offered reductive accounts which he thinks are plausible, meaning that he thinks it plausible that moral facts exist and (presumably) that our moral judgments are justified. Indeed, he claims explicitly that ‘there is empirical evidence that there are (relational) moral facts’ (1977: 132).

Let us use Harman’s own example to discuss the issue. He asks us to imagine that ‘you see some children pour gasoline on a cat and ignite it’ (1977: 4). He continues:

[Y]ou make a moral judgment immediately and without conscious reasoning, say, that the children are wrong to set the cat on fire. ... In order to explain your making [this judgment], it would be reasonable to assume, perhaps, that the children really are pouring gasoline on a cat and you are seeing them do it. But [there is no] obvious reason to assume anything about “moral facts,” such as that it is really wrong to set the cat on fire. ... Indeed, an assumption about moral facts would seem to be totally irrelevant to the explanation of your making the judgment you make. It would seem that all we need assume is that you have certain more or less well articulated moral principles that are reflected in the judgments you make, based on your moral sensibility. It seems to be completely irrelevant to our explanation whether your intuitive immediate judgment is true or false. (1977: 7)

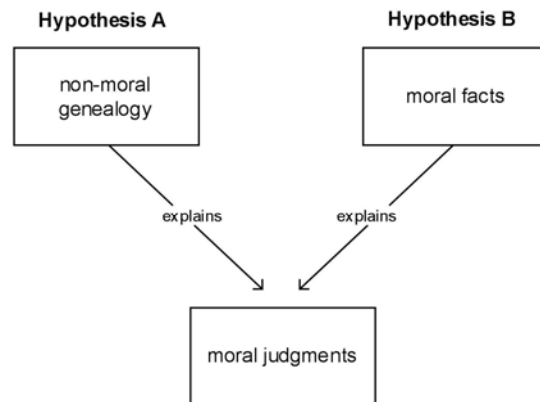
Imagine that Jane judges a particular episode of cat-burning wrong. Suppose we could explain the event of Jane’s judgment entirely and adequately in the terms of physics and chemistry—an explanation in which the words ‘cat’, ‘burning’, and ‘wrongness’ do not even appear. Does this show that burning cats play no part in explaining Jane’s judgment? No, for burning cats can be *reduced* to physics and chemistry, allowing us to recognize that the burning cat was implicitly present (so to speak) in our causal explanation of her judgment. But is there any comparable story that we can tell to explain how *wrongness* fits into the same naturalistic world? Harman’s point is that if we have a complete naturalistic explanation for why it *seems* to Jane that the action is wrong, but have no clear idea of how wrongness fits into (reduces to) this explanation, then the actual existence of wrongness isn’t needed to explain anything in the situation, in which case we have no reason to believe that it is a part of this cat-burning episode at all. And, obviously, this reasoning is supposed to be entirely generalizable. Whenever we judge something morally wrong there is always a complete explanation of the judgment that neither presupposes moral facts nor acts as a reductive base of moral facts. With moral judgments thus explained without recourse to moral facts, and in the absence of anything else whose explanation requires us to posit moral facts (for what phenomenon could there be for which our tendency to seek a moral explanation does not ultimately depend on our having made a moral judgment?), we have no reason to believe that anything at all is morally wrong. (The same holds for other moral qualities.)

Harman is fairly vague about what he takes an adequate ‘reduction’ to be. He says that it ‘need not involve definitions of a serious sort’, and he gives as an example the relation between tables and clusters of atoms (1985: 33). This lines up with my above claim that burning cats ‘reduce to’ physics and chemistry, even though there are no semantic relations permitting deductions between propositions containing the words ‘cat’ and ‘burning’ and propositions

expressed in the language of physics and chemistry. Nevertheless, we at least think we understand how burning cats fit into the world as described by physics and chemistry—we have a story to tell explaining how a burning cat is a physical, chemical entity. Much more needs to be said about this notion of *reduction*, of course, but all we need note here is that it is a very broad notion, apparently covering many positions that would ordinarily be categorized as non-reductionist. Harman’s well-known opponent in this debate, Nicholas Sturgeon, uses a much narrower conception of *reduction*—one that concerns ‘whether moral explanations can be given reductive definitions in some nonmoral vocabulary’ (1985: 59; see also his 1986). (In my opinion, this terminological discrepancy explains a disappointingly large proportion of the debate between Harman and Sturgeon.) In what follows I will go along with Harman’s broad usage, for there is no rationale in his argument requiring that the moral facts be describable in the language of the natural sciences; a purely ontological relation will suffice.

Figure 1 illustrates two competing hypotheses concerning how to explain the phenomenon of moral judgments.

Figure 1

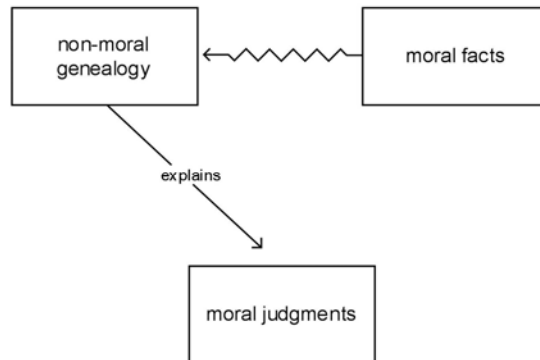


Remember that hypothesis A is supposed to be empirically confirmed. Overlooking this may encourage one to consider this argument as unimpressively analogous to standard challenges from the philosophical skeptic. For example, you may believe that right now you are sitting reading a book, but an annoying skeptic can always provide an alternative hypothesis (‘hypothesis A’), consistent with all the data available to you, according to which you aren’t reading a book at all. Perhaps you are really a brain floating in a vat of nutrients and being fed neural impulses corresponding to sitting reading a book. But the view under discussion here does not come so cheap. It is not just that in this case we can *make up* a consistent hypothesis according to which a bunch of our everyday beliefs are false; rather, we might have empirical evidence supporting the hypothesis that explains how these beliefs came about but does not require that they be true. The argument does not depend on invoking extreme standards for epistemic justification; the skeptic is not requiring people to consider outlandish brain-in-vat-type possibilities that they would ordinarily scoff at.

Given figure 1, one might have thought that hypothesis B should be excised from the picture with a swift slash from Ockham’s Razor, since we have a complete explanation of moral judgment with no need to posit any extra ontology in the form of moral facts. This appears to be

how Michael Ruse argues for the conclusion that the evolutionary basis of morality undermines morality when he claims that ‘the objective foundation for morality is redundant’ (1986: 254). This, however, is too hasty, as figure 2 shows.

Figure 2



The crooked line indicates a reduction relation (in the broad ontological sense of the term). If the moral facts are reducible to the non-moral facts invoked in the genealogical explanation, then the former cannot be eliminated on grounds of parsimony, any more than cats should be eliminated from our ontology because we can explain them in terms of physics. Let us give the label ‘moral naturalism’ to the view that such a relation holds between moral properties and naturalistic properties.⁷ Thus it is moral naturalism that promises to answer Harman’s challenge.

Something about which Harman is correctly adamant is that acknowledging the mere possibility of moral naturalism saving the day accomplishes next to nothing if it is not backed up with a concrete theory explaining how the moral fits into the natural world (unless we have some really compelling reason for assuming that such a theory is forthcoming). Nor is it enough merely to offer a naturalistic reduction, for doing so is easy. Charles Stevenson (1937: 14) once suggested that ‘something is morally good if and only if it is pink with yellow trimmings’. (He was making a philosophical point, not putting forward a serious contender!) When a moral naturalist offers a theory, ‘we have to be able to believe in this account’ (Harman 1986: 63); the account must satisfy our criteria of adequacy. In the following section I will highlight the extent to which this matter is empirical.

Moral naturalism

Most people who doubt moral naturalism do so because they judge moral properties to essentially have some kind of feature that natural properties do not (or cannot) have. For example, many moral philosophers have believed that moral facts must have some special kind of inescapable practical authority (let’s call it ‘practical clout’) that the world as described by science just cannot supply (see Mackie 1977; Joyce 2001). Though I am sympathetic to this view, it is not my intention on this occasion to endorse it; rather, in what follows I will use the phrase ‘practical clout’ as an example of—almost as a placeholder for—whatever attribute the anti-naturalist judges to be both essential to morality and unsatisfied by the natural world; so it should be borne

in mind that the anti-naturalist argument could conceivably run while focused on a different problematic property altogether. (Thus I feel free to sidestep the delicate task of specifying in any detail what this special kind of inescapable authority amounts to.⁸)

In response to this kind of anti-naturalist conviction, the naturalist can go either of two ways. (i) She can attempt to show that this practical clout can be accommodated within a naturalistic framework, or (ii) she can deny that it is really a requirement of morality at all. It is this second strategy that I will discuss here (noting for the record that I believe the prospects of the former strategy are dim). Naturalists of the latter stripe sometimes assume that the notion of *practical clout* is nothing more than an extravagance dreamed up by a philosopher, while others acknowledge that it is usually thought of as important to morality (that its satisfaction is a desideratum) but judge that a moral theory satisfying enough of our other moral platitudes may be close enough to count as an acceptable naturalization. Obviously, to assess this properly we would need to have a much more careful understanding of what practical clout amounts to. But even given this knowledge we would face a further question: How do we know whether a feature (e.g., clout) is *essential* to moral discourse? How do we know when an offering that satisfies some but not all of our pre-theoretical desiderata counts as ‘close enough’ to deserve the name ‘*moral naturalism*’ (as opposed to amounting to a naturalization of some properties that are similar to, but distinct from, moral properties)?

Philosophers have no settled views on how to adjudicate or even conceptualize such a debate. If one person asserts that something *is* a non-negotiable feature of some concept and another person denies this, where should they take their dispute? David Lewis makes use of the distinction between speaking strictly and speaking loosely: ‘Strictly speaking, Mackie is right: genuine values would have to meet an impossible condition, so it is an error to think there are any. Loosely speaking, the name may go to a claimant that deserves it imperfectly. ... What to make of the situation is mainly a matter of temperament’ ((1989) 2000: 93). Although this is unobjectionable up to a point, it really just postpones the problem, for we might continue to argue about whether something exists *even when we have confined ourselves to speaking loosely*. Presumably we will not accept a theory that allows that ghosts and witches exist—even loosely speaking—and certainly we don’t want one that tolerates that the matter might be settled according to ‘temperament’.

The absence of an obvious answer ensures that this is the resting place of many a metaethical debate. My own view is that what determines the answer to such matters is how the relevant population of speakers would collectively decide. Sometimes discoveries lead us to decide that a concept (e.g., *phlogiston* or *witch*) is hopeless; sometimes we prefer to revise the concept, extirpate the problematic element, and carry on much as before (e.g., the concept *simultaneity* survived the discovery that it’s all relative; the concept *polymer* survived the discovery that they are macromolecules rather than colloids). But there is absolutely no reason to assume that when groups of humans make such decisions they are following a hidden principle. Who is to say that our collective decisions on such matters aren’t influenced by the most trivial of things, such as advertising jingles or the way a word is used in a popular movie? This reveals that in many cases there really may be no fact of the matter as to whether something is an essential feature of a concept. Perhaps given one cultural milieu we’d decide in the positive, but given a somewhat different (but not remarkably different) cultural setting we’d decide in the negative. In light of

this, there seems something hopeless about a lone philosopher asserting with confidence that some disputed attribute is or is not an a priori requirement of a concept, or declaring that some imperfect satisfier of the platitudes surrounding a concept is or is not ‘close enough’ to count as a revised continuer of the original flawed concept. At best, such assertions could be hypothetical predictions about what a group of humans in circumstances like ours would probably do if forced to make a choice—predictions that are, obviously, burdened with difficulty and more often than not untestable in practice.

But perhaps all is not lost. We might stand a decent chance of getting at the answer if we give consideration to what the concept is *used for*, what practices it undergirds, and then ask whether a revised concept, with the problematic element discarded, could carry on playing that role. For example, even when we realized that nothing is absolutely simultaneous with anything else, the relativistic notion of simultaneity was able to take over seamlessly, since it works just as well in everyday contexts for creatures whose movements don’t approach a significant fraction of the speed of light. We can *use* the concept of relative simultaneity in the same way as we can use absolute simultaneity, which suggests that the change didn’t amount to replacing one concept with a different concept at all, but rather we just made a revision internal to a single concept. Thus we are not forced to the radical position that every pre-Einsteinian assertion of two events occurring simultaneously is false. By comparison, when we discovered that there are no diabolical supernatural forces in the universe, we had no further use for the concept *witch*. Perhaps we could have carried on applying the word ‘witch’ to women who play a certain kind of local cultural role on the margins of formal society—perhaps we might even have located a cluster of naturalistic properties that all and only these women have—but carrying on in this way would not have allowed us to *use* the word ‘witch’ for the purposes to which we had previously put it: to condemn these women for their evil magical influence and justify their being killed. Thus, there was little point in persisting in using the word ‘witch’ to stand for certain instantiated naturalistic properties; we dropped it and concluded that all historical assertions that certain women were witches—even the loosely spoken ones—were false.

The question that needs to be asked, therefore, is whether moral discourse could carry on playing whatever role(s) it does play when practical clout (or whatever other problematic attribute is under consideration) is eliminated. And this, I want to point out, is largely an empirical matter. First, we need to know what roles moral thinking and moral discourse do in fact play, both for the individual and the group. It is something of a travesty in moral philosophy that this question has so rarely been squarely addressed; philosophers have largely contented themselves with the unexamined assumption that morality is a Good Thing without which we’d all be worse off. But intuitions on this matter, even widespread ones, will not stand in for concrete data. Most conspicuously, we need to know the connection between moral thinking and motivation: Does thinking of an action as ‘obligatory’ (say) really strengthen one’s resolve to perform it (or to punish others who omit to perform it)? Are people prone to guilt really less likely to transgress against norms? Does a society where moral norms are publicly stressed (playing an explicit role in education, say) enjoy less criminal activity? Do moral convictions make interpersonal and/or intergroup conflicts more intractable or more violent? Once we have a reasonable set of answers to such a posteriori questions we need to perform a second task: assessing whether a discourse could carry on playing these roles with practical clout (or

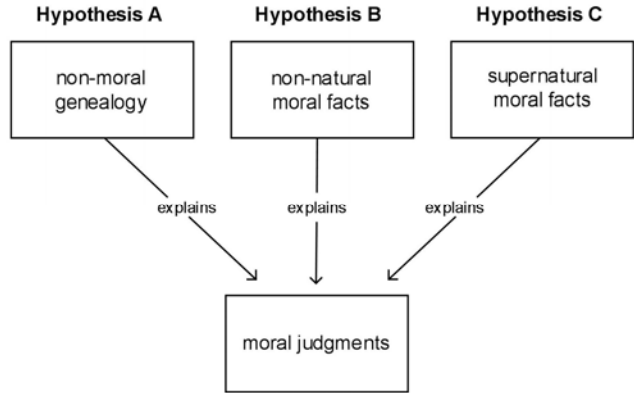
whatever) extirpated.⁹ This task may involve examining persons with pathologies that leave them unable to employ practical clout in guiding their deliberations. (Are such people also chronically weak of will, transgressive, unhappy, or uncaring?) The task may involve seeking societies or sub-societies for whose members the idea of practical clout plays no role. There are many ways to approach the question, and I don't want to deny that sitting in one's armchair and speculating may be among them (such activity often casts up promising ideas)—but the ultimate arbiter must be the body of a posteriori data issuing from such disciplines as psychology, anthropology, and experimental economics. Even an evolutionary genealogy may shed light on the subject, since an understanding of how having a moral sense advanced the reproductive fitness of our ancestors may have a lot to do with what role(s) it continues to play. To the extent that the evidence may support a hypothesis according to which moral judgment advanced ancestral fitness precisely because it was imbued with a special kind of no-questions-asked authoritative clout (see Ruse 1986; Dennett 1995; Joyce 2001, 2006), the position of anyone claiming that morality lacks such authority, or can carry on just fine without such authority, is weakened.

Harman's challenge again

Many philosophers are global naturalists (see note 7). Unless they plump for some kind of noncognitivism, the only metaethical options they recognize are either some form of moral naturalism or moral nihilism. Thus, for them, any argument that defeats moral naturalism (assuming noncognitivism is out of the picture) will have succeeded in showing that there are no moral facts at all. Note that references to Harman's challenge and to the genealogy of moral judgments will have played no part in establishing any such moral nihilism, except, perhaps, for providing an explanation of why we all have been so systematically misguided in believing in morality for all this time.

But many philosophers are not global naturalists. *Moral non-naturalists* hold that moral properties exist but enjoy a certain autonomy from the world as described by the natural sciences: They are not identical to, reducible to, or supervenient upon any natural properties. *Moral supernaturalists* hold that moral properties exist but depend for their existence on some kind of supernatural phenomenon—most obviously, God's will or commands. Even if these stances are not abundantly populated by contemporary philosophers, I think it is safe to say that they come the closest to capturing what ordinary speakers believe. Thus Harman's challenge still has real work to do. Consider figure 3.

Figure 3



The challenge, remember, is that hypothesis A promises to explain all our moral judgments, leaving us without need to posit any moral facts (i.e., with no reason to assume that any of our moral judgments are true) unless the moral facts are somehow implicitly buried in hypothesis A. The only way that moral facts could be implicitly buried in a scientific genealogical hypothesis is if some kind of moral naturalism were true. (See figure 2.) Thus, if we have reason to doubt moral naturalism, we will be left with figure 3, and this time Ockham’s Razor really can come in and do its thing, for non-naturalism and supernaturalism do posit extra ontology in the world, but the presence of the non-moral genealogy (hypothesis A) shows this ontology to be explanatorily superfluous. Hypotheses B and C can be excised.

I am spelling out this dialectic carefully because it is often, to my mind, misunderstood. Harman’s challenge is often seen as a problem for moral naturalism. It does represent a demand that the naturalist articulate her case clearly, showing how the moral connects to the natural—but if the case can be so articulated, the challenge fades (in fact it is hard to imagine that any naturalist might have supposed that she could present his theory persuasively without passing the challenge). And to the extent that the naturalist can make her case, non-naturalism and supernaturalism become less plausible. But if the naturalist *cannot* make her case, Harman’s challenge seems to make non-naturalism and supernaturalism obsolete. In other words, once we have a complete non-moral genealogy of moral judgment, if moral naturalism succeeds non-naturalism and supernaturalism are sunk, and if moral naturalism fails non-naturalism and supernaturalism are sunk. Thus non-naturalism and supernaturalism suffer most in this argumentative fray, whereas the moral naturalist is defeated only through independent arguments having nothing in particular to do with Harman’s challenge

The conclusion of the argument outlined in this paper is not moral nihilism (the view, roughly, that all our moral judgments are false). Pointing out that we have no reason to believe in moral facts does not imply that we have reason to *disbelieve* in them. I have no reason to believe that the number of hairs on my head is an odd number (even correcting for vagueness), but I’d be foolish to conclude that I therefore have reason to disbelieve that it’s an odd number, for that would be a reason to believe that it’s an even number. Though I am not justified in holding that the number of hairs on my head is odd, it *could* be odd (in fact, there is a pretty decent chance that it *is* odd). I have outlined an argument according to which the status of morality may be analogous. Assuming that there are such things as moral beliefs, if we have an empirically confirmed theory about where these moral beliefs come from, and if this theory doesn’t state or

imply that they're true, and doesn't hold as a background assumption that they are true, and their truth is not surreptitiously buried in the theory by virtue of any form of moral naturalism, then this amounts to the discovery that our moral beliefs are products of a process that is entirely independent of their truth, which forces the recognition that we have no grounds one way or the other for maintaining these beliefs. There are a lot of 'if's in the foregoing reasoning, of course, but none of them are particularly far-fetched. The principal object of this paper has been to draw attention to the number of 'if's that can be settled only by empirical inquiry

Note that this conclusion will not be swept aside by an appeal to any of the standard epistemological theories. Consider, first, epistemological conservatism, which holds that one may be justified in maintaining a belief, even in the absence of any positive evidence, simply because held beliefs have a presumption of rationality (they are 'innocent until proven guilty'). However, any plausible version of conservatism will accommodate the common-sense view that there are certain attributes that a belief may have that cast it into doubt, such as being clouded by emotion or subject to illusion (see Sinnott-Armstrong 2005). The provision of an empirically confirmed theory that explains the origin of certain beliefs but which at no point implies or presupposes their truth is one such defeater of presumed justification. Consider, second, process reliabilism, which holds that a belief is justified if and only if it is the product of a process that reliably links beliefs with truth. However, the presence of an empirically confirmed theory that explains the origin of our moral beliefs but which at no point implies or presupposes their truth, constitutes evidence that these beliefs flow from an unreliable process. Consider, third, epistemological coherentism, which holds that a belief is justified if and only if it sits comfortably (in some sense to be specified) with other held beliefs. Suppose we start with the belief that the Nazis were evil; but suppose we add to our doxastic set the belief that humans have beliefs concerning evil due to the circumstances of the social life of our ancestors, and this genealogy nowhere presupposes the truth of such beliefs. Suppose further that we endorse certain epistemological platitudes about when held beliefs are defeated or rendered dubious, such as the truism that independent confirmation is needed for a belief when it arises from a process that nowhere assumes its truth. The most coherent way of putting all these beliefs together would be to conclude that the moral belief stands in need of some explicit confirmation; thus doubt is cast upon the original moral belief *by the coherentist's own lights*. These brief comments are, of course, headlines rather than arguments, but I contend that, quite generally, on no epistemological theory worth its salt should the justificatory status of a belief remain unaffected by the discovery of an empirically-supported theory that provides a complete explanation of why we have that belief while nowhere assuming its truth.

NOTES

Much of this paper is a condensed version of arguments presented in *The Evolution of Morality* (MIT Press, 2006). Many passages are taken straight from this book.

1. It should be noted that even moral naturalists who *are* interested in issuing ‘ought’ claims are, these days, unfazed by the injunction against deriving ‘ought’ from ‘is’. The objection fails simply because it is not incumbent on the moral naturalist to offer a deductively valid argument whose conclusion contains the word ‘ought’ but none of whose premises contain this word. One can claim that moral properties are identical to (or supervene upon) natural properties, while denying that there is a *semantic* or *deductive* relation between propositions about the natural world and propositions involving normative language.
2. I admit the possibility that this fanciful example may not even be a conceptual possibility, but I suggest that on occasion even impossible thought experiments may serve a useful pedagogic or intuition-priming role.
3. According to one family of epistemological theories, knowledge of the genealogy can *render* our beliefs unjustified (i.e., our beliefs were justified prior to gaining this knowledge); according to another family, this knowledge reveals that the beliefs were unjustified all along. Since I’m not taking sides on this matter, my conclusion is ambiguous in this way.
4. Some philosophers doubt that mathematical propositions are true. But the dialectic within which I am working here assumes that if an argument that moral beliefs are unjustified or false would by the same logic show that believing that $1 + 1 = 2$ is unjustified or false, this would count as a *reductio ad absurdum*.
5. To pinch a phrase from Crispin Wright 1992: 162.
6. Certain psychologists use the word ‘emotivism’ differently from its philosophical usage. See Joyce forthcoming *a*, for discussion of this point.
7. For a fuller definition, see Copp 2004. Moral naturalism is not to be confused with *global naturalism*, a more general stance according to which the only kinds of things whose existence we ought to countenance are things that fit into a unified scientific framework. The global naturalist will want to ‘naturalize morality’ in the sense of providing a scientifically respectable account of moral institutions and practices, of moral psychology, and of moral genealogy. But this doesn’t amount to being a moral naturalist in the sense outlined, since it is consistent with holding that there exist no moral facts at all.
8. On other occasions I have not so shamelessly shirked this task: see Joyce 2001, 2006. For a careful and useful analysis of the concepts *inescapability* and *authority*, see Brink 1997.
9. Note that the task cannot be conceptualized as seeking *a moral system* that lacks clout, for whether something lacking clout deserves the name ‘moral’ is exactly what we are endeavoring to decide. What we would be looking for, I take it, is some kind of ‘normative system’—a set of values and/or imperatives that guide action—within which clout has no place.

REFERENCES

- ALEXANDER, R. 1987. *The biology of moral systems*. New York: Aldine de Gruyter.
BLACKBURN, S. 1984. *Spreading the word*. Oxford: Oxford University Press.

- BLACKBURN, S. 1993. *Essays in quasi-realism*. Oxford: Oxford University Press.
- BRINK, D. 1997. Kantian rationalism: Inescapability, authority, and supremacy. In *Ethics and practical reason*, edited by G. Cullity and B. Gaut. Oxford: Oxford University Press.
- BRUENING, W.H. 1971. G.E. Moore and 'is-ought'. *Ethics* 81: 143-149.
- BUTTERWORTH, B. 1999. *What counts: How every brain is hardwired for math*. New York: The Free Press.
- CAMPBELL, R. 1996. Can biology make ethics objective? *Biology and Philosophy* 11: 21-31.
- COPP, D. 2001. Realist-Expressivism: A neglected option for moral realism. *Social Philosophy and Policy* 18: 1-43.
- COPP, D. 2004. Moral naturalism and three grades of normativity. In *Normativity and Naturalism*, edited by P. Schaber. Frankfurt: Ontos-Verlag.
- DARWIN, C. (1879) 2004. *The descent of man, and selection in relation to sex*. London: Penguin Books.
- DENNETT, D.C. 1995. *Darwin's dangerous idea*. New York: Simon and Schuster.
- 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: 2105-2108.
- HAIDT, J. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108: 814-834.
- HAIDT, J. and C. JOSEPH. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus* 133: 55-66.
- HARMAN, G. 1977. *The nature of morality: An introduction to ethics*. New York: Oxford University Press.
- HARMAN, G. 1985. Is there a single true morality? In *Morality, reason and truth*, edited by D. Copp & D. Zimmerman. New Jersey: Rowman & Allanheld.
- HARMAN, G. 1986. Moral explanations of natural facts: Can moral claims be tested against moral reality? In *Spindel conference: Moral realism*, edited by N. Gillespie (*The Southern Journal of Philosophy* suppl. vol. 24).
- JOYCE, R. 2001. *The myth of morality*. Cambridge: Cambridge University Press.
- JOYCE, R. 2006. *The evolution of morality*. Cambridge, MA: MIT Press.
- JOYCE, R. Forthcoming a. What neuroscience can (and cannot) contribute to metaethics. *Moral psychology: Morals in the brain*, edited by W. Sinnott-Armstrong.
- JOYCE, R. Forthcoming b. Expressivism, motivation internalism, and Hume. In *Reason, motivation, and virtue*, edited by C. Pigden. University of Rochester Press.
- KATZ, L.D. (editor), 2001. *Evolutionary origins of morality: Cross-disciplinary perspectives*. Thorverton, UK: Imprint Academic.
- KITCHER, P. 1998. Psychological altruism, evolutionary origins, and moral rules. *Philosophical Studies* 89: 283-316.
- KITCHER, P. 2005. Biology and ethics. In *The Oxford handbook of ethics*, edited by D. Copp. New York: Oxford University Press.
- LEWIS, D.K. (1989) 2000. Dispositional theories of value. In his *Papers in ethics and social philosophy*. Cambridge: Cambridge University Press.
- MACKIE, J.L. 1977. *Ethics: Inventing right and wrong*. New York: Penguin Books.

- MOLL, J., R. DE OLIVEIRA-SOUZA, and P.J. ESLINGER. 2003. Morals and the human brain: A working model. *NeuroReport* 14: 299-305.
- RAILTON, P. 2000. Darwinian building blocks. In *Evolutionary origins of morality: Cross-disciplinary perspectives*, edited by L.D. Katz. Thorverton, UK: Imprint Academic.
- RUSE, M. 1986. *Taking Darwin seriously*. Oxford: Basil Blackwell.
- SINNOTT-ARMSTRONG, W. 2005. Moral intuitionism meets empirical psychology. In *Metaethics after Moore*, edited by T. Horgan and M. Timmons. New York: Oxford University Press.
- STEVENSON, C.L. 1937. The emotive meaning of ethical terms. *Mind* 46: 14-31.
- STURGEON, N.L. 1985. Moral explanations. In *Morality, reason and truth*, edited by D. Copp and D. Zimmerman. New Jersey: Rowman and Allanheld.
- STURGEON, N.L. 1986. Harman on moral explanations of natural facts. In *Spindel conference: Moral realism*, edited by N. Gillespie (*The Southern Journal of Philosophy* suppl. vol. 24).
- WRIGHT, C. 1992. *Truth and objectivity*. Cambridge, MA: Harvard University Press.