Expressivism, Motivation Internalism, and Hume

Richard Joyce

Penultimate draft of the paper in C. Pigden (ed.), *Hume on Motivation and Virtue* (Palgrave MacMillan, 2010).

David Hume is often taken to be a moral expressivist (Flew, 1963; Ayer, 1980, pp. 84-5; Price, 1988, p. 6; Snare, 1991; Harman, 1996, p. 97). He is, moreover, often taken to have presented in the *Treatise* one of the strongest arguments for moral expressivism: the so-called *motivation argument*. As a metaethicist, I am interested in whether expressivism is true, and thus interested in whether the argument that people think they find in Hume is a sound one. Not being a Hume scholar (but merely a devoted fan), I am less interested in whether Hume really was an expressivist or whether he really did present an argument in its favor. Hume's metaethical views are very difficult to nail down, and by a careful selection of quotes one can present him as advocating expressivism, or cognitivist subjectivism, or moral skepticism, or a dispositional theory, or an ideal observer theory, or even utilitarianism. It is entirely possible that Hume's position is indeterminate when considered against these terms of modern moral philosophy; it is also entirely possible that he was hopelessly confused (much as it pains me to admit it). However, I doubt very much that Hume should be interpreted as an expressivist in any straightforward manner, and therefore I am doubtful that he should be interpreted as arguing in its favor. Most of this paper does not discuss Hume directly at all: I critically discuss the motivation argument and I advocate a certain positive metaethical view—one that mixes elements of traditional expressivism with elements of cognitivism. This position is neutral between moral realism and radical moral skepticism. I close by wondering—very briefly—whether Hume might have held such a view. Given my reservations about the determinacy of Hume's metaethical outlook, the case is not pressed with any vigor, but because it is an interpretation of Hume that has not, so far as I know, been articulated before, it may be of interest to note that it seems to be consistent with much of what he says—at least as much as any other precise interpretation.

I. Expressivism and motivation internalism

Let me start by clarifying terminology. *Noncognitivism* is the metaethical view according to which public moral judgments do not express beliefs (that is, are not assertions) in spite of the fact that they are typically formed in the indicative mood. Thus defined, noncognitivism is a view of what moral judgments *are not*—leaving open space for many different forms of noncognitivism claiming what moral judgments *are*. One positive form, *prescriptivism*, holds that moral judgments are really commands.[1] My focus in this paper is on another form, *expressivism*, which holds that moral judgments function to express desires, emotions, pro-/con-attitudes, or (in Simon Blackburn's words) 'a stance, or conative state or pressure on choice and action' (1993, p. 168). I treat 'expressivism' and 'emotivism,' as they appear in metaethical discussions, as synonyms.

Why might one be tempted by noncognitivism (and expressivism in particular)? First, noncognitivism sidesteps a number of thorny metaethical puzzles that face the cognitivist. The cognitivist thinks that when we make a public moral judgment, like 'That act of stealing was wrong,' we are asserting that the act of stealing in question instantiates a certain

[1] See, for example, Carnap, 1935; Stevenson, 1937.

property: wrongness. But queries arise: What kind of property is wrongness? How does it relate to the natural properties instantiated by the action? How do we have epistemic access to the property? How do we confirm whether something does or does not instantiate the property? The difficulty of answering such questions may lead one to reject the presupposition that prompted them: One might deny that in making a moral judgment we are engaging in the assignment of properties at all. Such a rejection, roughly speaking, is the noncognitivist proposal. Second, the noncognitivist might claim the advantage of more readily accounting for certain aspects of moral disagreement—for example, its vehemence and intractability (see Stevenson, 1963, essays 1 and 2). The third traditional consideration in favor of noncognitivism is the subject of our attention: that noncognitivism does a better job than its rival of explaining the apparent motivational efficacy of moral judgment (see Smith, 1994, chapters 1-2).

Those who advocate this third argument for noncognitivism often look to Hume for a precedent, finding solace especially in the following passage:

> Since morals, therefore, have an influence on the actions and affections, it follows, that they cannot be deriv'd from reason; and that because reason alone, as we have already prov'd, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason. (*T* 3.1.1.6/457)

It is not unusual to find Hume's premise that 'morals excite passions' formulated as *motivation internalism*:

> SIMPLE MOTIVATION INTERNALISM (simple-MI):
> It is necessary and *a priori* that anyone who judges that she is morally required to ϕ will be (defeasibly) motivated to comply.

My object here is not to assess the truth of simple-MI, but to investigate its logical relation to metaethical expressivism. There are broadly two ways that one might use simple-MI in favor of expressivism. One might straightforwardly think that simple-MI entails expressivism, presumably with the supplement of some unobjectionable additional premise(s). Alternatively, one might think that simple-MI is a desideratum that any metaethical theory must strive to satisfy, and thus if expressivism were to entail simple-MI this would count very much in expressivism's favor. Let us examine both these implications in turn.

Does expressivism imply simple motivation internalism?

No, it does not. Expressivism, as I have stated it, is a thesis about what kind of mental states are expressed by moral judgments. It is vital to note that the notion of *expression* that is relevant here is non-causal. One can express a mental state while not having that state, and perhaps having never had that state. Consider a promise, which is a kind of speech act by which we express intentions or commitments. If I, in unexceptional circumstances, say to you 'I promise to be at the party tonight,' then I have expressed an intention to be at the party tonight. But my promise may nevertheless be insincere, in the sense that I have no intention of coming to the party, and have never had any intention to come. Since that intention appears nowhere in my mental repertoire, it *cannot be the cause* of my promise utterance. Now consider the following thesis:

SIMPLE PROMISING INTERNALISM:

It is necessary and *a priori* that anyone who promises to ϕ (thereby expressing the intention to ϕ) has the intention to ϕ.

It is clear that simple promising internalism does not follow from the thesis that promises express intentions. The phenomenon of insincerity is sufficient to demonstrate this. If, then, we construe metaethical expressivism as a thesis about what kind of speech act moral judgments are—which is natural if we read it as the denial that moral judgments are assertions, since *assertion* is a category of speech act (see Austin, 1962; Searle, 1969)—then the phenomenon of insincerity is entirely sufficient to show that simple-MI does not follow.

The reader might be wondering precisely what kind of relation is denoted by 'expression' in this context if it is not a causal one. This is something I discuss later. One might also object that I have construed either expressivism or motivation internalism (or both) incorrectly. They are, I admit, both theses for which there is disagreement as to their correct formulation. I will consider variants in due course, but first let me consider the reverse implication with these simple formulations.

Does simple motivation internalism imply expressivism?

No, it does not. From the fact that there is a necessary (and *a priori*) connection between a kind of mental state and a kind of speech act, it does not follow that the speech act *expresses* that mental state. Let us consider promises again, and consider what criteria must be satisfied in order for X to succeed in making a promise (albeit possibly an insincere one) to Y. The most complete answer to this question comes from John Searle (1969, pp. 57-61), who painstakingly delineates the conditions that must obtain if *S* is to promise that *p* to *H* via uttering *T*. I shall not rehearse all Searle's items, but just focus on a couple. First: *H would prefer S's doing A to his not doing A,*[2] *and S believes H would prefer his doing A to his not doing A.* Second: *It is not obvious to both S and H that S will do A in the normal course of events*. Both these criteria make essential reference to the parties to the promise having certain *beliefs*. These connections are necessary and *a priori*: It is not possible that any person could succeed in making a promise to another person without their having these beliefs. Yet we would hardly say that the act of promising *functions to express the belief that the promisee would prefer that the promised action be performed to its not being performed*—rather, the mental state expressed by the promise is as we originally stated: an intention or commitment. 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.

A related confusion has cropped up in some quarters over what metaethical conclusions might be drawn from certain recent empirical results that show the important role that emotions play in moral judgment (see Greene et al., 2001; Greene & Haidt, 2002; Moll et al., 2002; Haidt, 2001). Though these scientists' conclusions are not uncontroversial, let us take them at their word when they assert that 'recent evidence suggests that moral judgment is more a matter of emotion and affective intuition than deliberative reasoning' (Greene & Haidt, 2002, p. 517). This conclusion is often referred to as 'emotivism' (Haidt, 2001, p. 816; Greene et al., 2001, p. 2107; Greene et al., 2004, p. 397). Anthropologist Daniel Fessler

[2] '*A*' denotes a future act that proposition *p* predicates of *S*.

claims that 'emotivist perspectives on moral reasoning hold that emotional reactions precede propositional reasoning' (Fessler et al., 2003, p. 31). Let us be a little bolder, and interpret this kind of emotivism as the claim that all moral judgments are *caused by* emotional arousal.

Clearly, this use of 'emotivism' among empirical scientists is very different from the metaethicist's usage, for whom it is usually treated as a synonym of 'expressivism.'[3] The terms 'emotivism' and 'expressivism' *in the metaethical tradition* do not denote a thesis about the causal origins of moral judgment; they denote (as we have seen) a thesis about what kind of mental state is expressed by public moral judgments. It might be best if we distinguish 'psychological emotivism' (the kind advocated by Jon Haidt, for example) from 'metaethical emotivism' (advocated by A.J. Ayer and Simon Blackburn, for example).[4] The crucial point is to note the logical independence of the two: Even if the evidence were to demonstrate that every single moral judgment is caused by emotional arousal (that is, demonstrate that psychological emotivism is true), this wouldn't imply anything about *the function of* moral language.

Consider, for example, any kind of metaethical theory according to which moral utterances are veiled reports about one's own mental states. According to this kind of cognitivist subjectivism, 'X is morally wrong' means 'I feel disapproval towards X.'[5] The latter is something that may be asserted, yet in the typical case it will have been prompted by emotional arousal in the speaker. Such a theory would be consistent with psychological emotivism but inconsistent with metaethical emotivism. On at least one occasion, Hume sounds like he endorses some such view: '[W]hen you pronounce any action or character to be vicious, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from the contemplation of it' (*T* 3.1.1.23/469). I doubt that this particular simplistic subjectivist interpretation of Hume ultimately withstands scrutiny, but I have little doubt that Hume should nevertheless be interpreted as some kind of psychological emotivist: He clearly and emphatically thinks that moral judgments have their origin not in the faculty of reason but in sentiment.[6] My point is that his advocacy of psychological emotivism does not commit him to the metaethical variety; there is very little evidence that he advocated, or even had much awareness of, metaethical emotivism/expressivism/noncognitivism. Many of his emotivist-sounding moments (for example, 'Morality … is more properly felt than judg'd of' [*T* 3.1.2.1/470]) may be smoothly

---

[3] See Joyce, 2008 for further discussion of this disparity between psychologists' and metaethicists' use of 'emotivism.'

[4] Páll Árdal (1966) once distinguished 'emotionism' from 'emotivism,' in a way that maps closely to the contrast between what I am calling 'psychological emotivism' and 'metaethical emotivism' (though perhaps it is even closer to the distinction that I will make shortly, between the mentalistic construal of expressivism and the metaethical construal of expressivism). I quite like Árdal's terminology, but it never took off. Jesse Prinz (2007) has recently reintroduced the term 'emotionism' for a somewhat different thesis.

[5] Charles Stevenson (1937; 1963) held a view of this sort, though he maintained that in addition the moral judgment includes a command.

[6] John Bricke writes of the subjectivist-sounding *Treatise* passage (*T* 3.1.1.23/469): 'The autobiographical rendering of evaluative sentences being so utterly implausible, it is fortunate that there is no reason whatever to think that Hume here means by 'meaning' what, when concerned with language, we now mean' (1996, p. 162). (See also Ayer, 1980, p. 84.) Nicholas Sturgeon (2008) interprets Hume as a subjectivist, but not the speaker-oriented kind mentioned here. After quoting the subjectivist-sounding passage, Sturgeon notes that Hume subsequently 'modifies this view to make the truth of one's ascription of virtue or vice depend, not on one's actual feelings, but on the feelings one would have under the right conditions, whether or not one now is (or even could be) in those conditions' (#). I am not sure whether Sturgeon's Hume ultimately counts as a psychological emotivist. This is my own fault, since I have characterized psychological emotivism only as carefully as is necessary to reveal its logical independence from metaethical emotivism—but have left it intentionally indeterminate in several respects.

interpreted as advocating psychological emotivism rather than metaethical. Indeed, in the earlier-quoted passage from the *Treatise* from which the motivation argument is drawn, what we in fact have, I would claim, is an argument for psychological emotivism, not metaethical emotivism.

In sum: There is clearly a significant difference between motivation internalism and psychological emotivism: one asserts a necessary connection between moral judgment and conative states whereas the other asserts a causal connection between the two. But advocates of either must be wary of the same potential pitfall: of assuming that their thesis implies or provides support for the thesis that metaethicists have called 'emotivism' or 'expressivism.' I have argued that expressivism does not imply motivation internalism, and nor does motivation internalism imply expressivism. The latter denial raises serious problems for an argument that some people have found in Hume: the motivation argument. Rather than claiming that Hume argued poorly, however, I urge the conclusion that it is a mistake to foist this argument for expressivism on him in the first place.[7]

Some will object to all of the preceding on the grounds that I have misconstrued motivation internalism, or misconstrued expressivism (or misconstrued both). So it is to variations on these theses that I now turn.

II. Variants of expressivism and motivation internalism

A prominent variation on motivation internalism is the following, due (inter alios) to Michael Smith (1994):

> NORMATIVE MOTIVATION INTERNALISM:
> It is necessary and *a priori* that anyone who judges that she is morally required to ϕ will be (defeasibly) motivated to comply, or she is irrational.

I mention this variant simply because it is well known, but in fact there is little to say about it here. The additional clause on the end, though important in other philosophical contexts, does not affect any of the arguments that I have already deployed to show the logical independence of expressivism and simple motivation internalism. This is less obviously so of the following variant:

> SINCERITY MOTIVATION INTERNALISM (sincerity-MI):
> It is necessary and *a priori* that anyone who sincerely judges that she is morally required to ϕ will be (defeasibly) motivated to comply.

(One may choose to add the suffix '…or she is irrational'; it doesn't matter to anything that follows.) The thesis is not always worded in just this way, but the term 'sincerely' is often included in statements of motivation internalism (see Hare, 1999, ch. 8; Timmons, 1999, p. 53; Svavarsdóttir, 2005, p. 186; Shafer-Landau, 2005, p. 142).

Sincerity-MI does not imply expressivism. Proof: Sincerity-MI is implied by simple-MI; therefore if sincerity-MI were to imply expressivism, then so too would simple-MI imply expressivism. But we have already seen that simple-MI does not imply expressivism,

---

[7] Rachel Cohen concurs that Hume's motivation argument 'is irrelevant to noncognitivism' (1996, p. 261); however, her positive interpretation differs from mine: she maintains that the argument concerns the nature of moral properties.

therefore nor does sincerity-MI. Thus, construing MI as sincerity-MI provides no succor for the (alleged) Humean motivation argument for expressivism.

But does expressivism imply sincerity-MI? One might be tempted to think so. Suppose for the sake of argument that expressivism is true: that when one (in ordinary circumstances) utters 'Stealing is morally wrong' (say), one thereby expresses some conative (that is, motivation-implicating) state. One might think that from this follows something about what it is for such a judgment to be sincere. One might think that any conation-expressing utterance of 'Stealing is morally wrong' is sincere if and only if the speaker actually has that conative state at the time of utterance. This would be an instance of a tempting general principle of speech-act sincerity, which I shall name after its advocate, John Searle (1969):

> SEARLE-SINCERITY:
> S's utterance U (at time *t*) is sincere iff U expresses mental state M, and S has M (at *t*)

With Searle-Sincerity as an additional premise, it does appear that expressivism implies sincerity-MI. The problem is that Searle-Sincerity, plausible as it may appear at a glance, is false.

On an earlier occasion (Joyce, 2002), I offered some counterexamples to Searle-Sincerity. I imagined someone saying 'Thanks!' as he left a dinner party in a distracted and hurried way, and claimed that though we might admit that at the moment of thanking he was feeling no gratitude whatsoever, nevertheless we wouldn't ordinarily call his utterance 'insincere.' A second counterexample along the same lines concerned an act of passing moral judgment. Michael Ridge (2006) criticizes these counterexamples on the grounds that they fail to take into account the fact that although the speaker may not have gratitude (say) as an *occurrent* emotion, he nonetheless may count as having that mental state (at the time of utterance) *dispositionally*. I harbor misgivings about dispositional mental states (especially emotional ones), but let us not pause to consider them now, for the main point is that Ridge nevertheless agrees with me that Searle-Sincerity is inadequate, and supplies counterexamples of his own that revolve around self-delusion. A person may believe himself to have mental state M when in fact he does not. If there is a speech act that expresses M, and the person performs that speech act, then it seems natural to say (Ridge argues) that the speech act is sincere, even though the speaker lacks the mental state that it expresses.

Ridge presents an alternative general thesis of speech-act sincerity:

> RIDGE-SINCERITY:
> S's utterance U (at time *t*) is sincere iff S believes that U expresses mental state M, and S believes that she has M (at *t*)[8]

I am concerned that Ridge's version of speech-act sincerity is also problematic, in that it presupposes that ordinary speakers have beliefs about a kind of expression-relation holding between utterances and mental states—but this, I suspect, is far too recherché a belief to require of ordinary speakers in order that they may be granted speech-act sincerity (even if we allow that the belief may be implicit, dispositional, and nonconscious). We have already seen the confusions that may entrap the unwary concerning causal versus conventional notions of *expression*; and if even analytic philosophers stumble over this, what hope should we have that ordinary speakers' beliefs are in order? Even children of a tender age have the

[8] For the sake of brevity I have stripped Ridge's thesis of a few details that don't matter on this occasion. For the full account see Ridge 2006, p. 501.

capacity to make assertions, ask questions, bark commands (and so forth)—and to do so in a sincere manner—yet surely they have no beliefs about what kind of mental states various utterances *express*. I once had an argument with a well-known philosopher (who shall remain nameless) who declared that metaethicists have no idea what they are talking about when they wonder about what mental states moral judgments *express*. He claimed that it was as if metaethicists had just assumed the existence of some mysterious relation holding between moral utterances and mental states, then given it a name—'expression' (though it may as well have been 'floog')—and have then expended endless energy arguing in circles about this baffling relation. This philosopher presumably (or at least conceivably) did *not* believe that any of his utterances *expressed* any mental states; he was sufficiently skeptical of the whole notion that he just withheld assent to such thoughts. Do we want to claim that this philosopher, despite himself, 'implicitly' had such beliefs? I wouldn't wish to maintain this; it seems a rejoinder of desperation. I might allow that his speech acts, despite his beliefs, *did* express various mental states, but I see little plausibility in the claim that, despite himself, he *believed* this fact. And yet, for all this, I am quite certain that this person was capable of making sincere assertions, sincere promises, sincere apologies, and so forth.

I am not here going to argue for an alternative general theory of speech-act sincerity, since I question the assumption that a general account is forthcoming or even particularly desirable. Perhaps what must be added to a promise to ensure its sincerity differs from what must be added to an assertion to ensure its sincerity, while both differ from congratulations, apologies, entreaties, thankings, and so on. If there is anything that unites these cases, in my opinion, it will revolve around the fact that insincere speech acts are ones by which the speaker attempts knowingly to mislead his/her audience—and such audience-directed intentions are not mentioned, nor entailed, by either Searle-Sincerity or Ridge-Sincerity. But I shall not develop this thought on this occasion, for the point that matters to our present purposes does not require it. For our present purposes I can even embrace Ridge-Sincerity. The point is that expressivism promises to imply sincerity-MI only with Searle-Sincerity as a bridging premise, but Searle-Sincerity has been refuted, and there is no reason to assume that whatever general thesis of speech-act sincerity replaces it (if there even is one) will also act as a bridge from one thesis to the other

Of course, if we have any account whatsoever of what a sincere moral judgment consists of, it will follows trivially that *some* kind of vaguely motivation-internalism-ish thesis will be implied by expressivism. Consider the generalized argumentative format…

PREMISE 1:	Moral judgments express conative state C (regarding the subject of the judgment)
PREMISE 2:	Sincere moral judgments have quality Q (vis-à-vis the mental state expressed)
THEREFORE:	If S sincerely judges that she is morally required to ϕ, then her judgment has quality Q vis-à-vis C (regarding her ϕing).

I call the conclusion 'motivation-internalism-ish' on the grounds that it asserts a relation between moral judgment and a kind of motivational state (C). (And if the two premises are necessary and *a priori* then the relation asserted in the conclusion will also be.) But in fact this conclusion is derived trivially, and has no useful role to play in the metaethical dialectic. Recall that the expressivist hope has been that some version of motivation internalism might have independent attractions, thus boosting the case for expressivism either by implying expressivism or by being revealed to be a desideratum that expressivism satisfies better than its rivals. But the kind of trivial MI-ish conclusion just mentioned can play neither role.

Another obvious problem with any version of motivation internalism that restricts itself to *sincere* moral judgments is that by encompassing only a proper subset of moral judgments, it fails to tell us anything about moral judgment *simpliciter*. By comparison, I could tell you something true about all moral judgments made on a Saturday—say, that they are 'weekend moral judgments'—but this obviously would tell us nothing about what we are interested in as metaethicists: namely, what a moral judgment *is*, what its necessary features are, and so on. Similarly, the fact that sincerity-MI reveals a connection between sincere moral judgments and motivational states doesn't imply any necessary connection between motivation and moral judgments *simpliciter*, since the motivational aspect may be smuggled in within the concept of *sincerity*. (If I am correct that the most promising account of speech-act sincerity will make reference to the speaker's intentions not to deceive, then sincerity will automatically bring motivation along for free, since these kinds of intentions are motivation-engaging states). Any version of internalism restricted to sincere moral judgments is compatible with the falsity of expressivism concerning moral judgments *simpliciter*.

I have lately been discussing variations on the thesis of motivation internalism, but it may also be objected that I have misconstrued the thesis of expressivism. One might, in particular, complain about my characterization of expressivism as a metaethical theory about *speech acts*; one might instead insist that expressivism is a theory about *mental states*: not about what kind of mental state moral judgments *express*, but about what kind of mental state moral judgments *are*. On such a view the applicability of the sincere/insincere distinction retreats, and, indeed, the whole troublesome expression relation conveniently evaporates.

This mentalistic construal of expressivism is unconventional. If we go back to the roots of noncognitivism is the early 20th century, we see pretty clearly that what is under discussion is the nature of moral *language*. In their influential 1923 book *The Meaning of Meaning*, C.K. Ogden and I.A. Richards speak of a use of the word 'good' which is 'purely emotive,' and '[w]hen so used the word stands for nothing whatsoever, and has no symbolic function' (p. 125). A.J. Ayer's noncognitivism was motivated by the question of how moral utterances might be meaningful *statements* ([1936] 1971).[9] Rudolf Carnap's noncognitivism was presented as the claim that a 'value statement' is not 'really an assertive proposition,' but is, rather, 'a command in misleading grammatical form' (1935, pp. 24-5). Charles Stevenson spoke of ethical judgments as having 'quasi-imperative force' which may be 'intensified by your tone of voice' (1937, p. 19). At no point in these classic works in the emergence of noncognitivism is it hinted that 'moral judgment' might be used primarily to denote a mental state.

But perhaps that is all a misguided historical idiosyncrasy, and perhaps we would do better now to treat moral judgments as a species of mental state. In this case, expressivism will be the theory that moral judgments are not beliefs, but rather some kind of conative state (to be specified). One obvious problem with such a decision is that it opens the possibility that moral judgments (qua mental states) might be conative while moral judgments (qua linguistic entities) might be assertoric. This is more or less the same possibility as was noted earlier, when I pointed out that metaethical cognitivism (interpreted in the orthodox manner) is compatible with either simple-MI or psychological emotivism. Of course, this observation

[9] Ayer thought that all meaningful statements must be either analytic or empirically verifiable. Given that moral utterances appear to be neither, Ayer was forced to claim that they are not meaningful statements. But rather than concluding that moral judgments are meaningless, Ayer's preferred conclusion is that they are not *statements*, but are, rather, ways of evincing ones emotions and issuing commands.

doesn't count as evidence against mentalistic expressivism, but rather indicates how confusing this way of characterizing theories might become.

In the present context, the important thing to note is that construing expressivism mentalistically would nullify the possibility of any argumentatively interesting relation holding between expressivism and motivation internalism. If expressivism is the theory that moral judgments *are* episodes of conative state C (where 'C' denotes something that is by stipulation necessarily motivation-engaging), then expressivism is essentially equivalent to the thesis of motivation internalism, which states that moral judgments necessarily engage motivation. The connections appear to be so trivial that arguing for either thesis by means of first establishing the other ceases to be a feasible dialectical strategy. This has particular relevance to the so-called motivation argument that is drawn from Hume's *Treatise*. From the premise that 'morals excite passions' (that is [putatively], that moral judgments necessarily engage motivations) one can certainly derive mentalistic expressivism—but only trivially: the conclusion essentially *is* the premise.[10,11]

One can read the preceding as presenting a dilemma to any expressivist tempted to employ the Humean motivation argument. Either expressivism is construed mentalistically (as a theory of what kind of mental states moral judgments *are*), in which case the argument is valid but question-begging, or it is construed linguistically (as a theory of what kind of mental states moral judgments *express*), in which case the argument is unsound. I have made clear my preference for construing expressivism in the latter fashion (thus impaling the advocate of the motivation argument on the second horn), and, in the course of discussion, I made much of the fact that the relevant expression relation must not be understood in a causal manner. I should like now to say more in a positive vein about how that expression relation ought to be understood. This discussion takes us well away from Hume—which is hardly surprising given my contention that thinking of Hume in the guise of a modern expressivist is a serious distortion—but I will close with some brief thoughts applying what we have learned to Hume.

III. Expressing mental states

When seeking explication of the sense in which types of speech acts express types of mental states, it is useful to start with Moore's Paradox. G.E. Moore (1942, p. 543[12]) noted the oddity of someone claiming:

1) I went to the pictures last Tuesday. But I don't believe that I did.

---

[10] It is possible that one might construe MI and mentalistic expressivism such that they have a different modal and/or epistemological status. MI, recall, is presented as a necessary and *a priori* thesis; perhaps mentalistic expressivism need not be. It is difficult, however, to see how this would create the possibility of a viable argumentative strategy from one to the other. From 'It is actually the case that X' we cannot conclude 'It is necessarily the case that X.' The reverse implication does hold, of course, but then the question is on what grounds we could establish the necessity claim as the antecedent. If we had any such grounds, then the consequent would hardly be in doubt.

[11] Frank Snare interprets the argument from motivation as aiming to establish mentalistic expressivism (he calls it 'emotivism'). He concludes that the influence of the argument is due to philosophers having 'been so completely convinced of the conclusion that they did not realize that the conclusion itself provides much of the reason for believing the premises' (Snare, 1975, p. 9).

[12] I am not sure whether Moore mentioned the paradox in print on any earlier occasion. Wittgenstein reports Moore having mentioned it in a lecture, probably from before World War 1. (Norman Malcolm mentions that Wittgenstein opined that this was the only work of Moore's that had ever 'greatly impressed him' [Malcolm, 1958, p. 66].)

It's called a 'paradox' because although it is not a logical contradiction (for it is perfectly possible that I went to the pictures last Tuesday while I don't believe that I did), to *state* the whole is to void the speech act of the first part, leaving the listener confused as to what should be assumed about the speaker's attitude towards his having been to the pictures last Tuesday. It makes (to quote J.L. Austin) 'a peculiar kind of nonsense' ([1970] 1990, p. 112). Moore presents the paradox using the category of assertion, but it seems we can find exactly the same phenomenon with other species of speech act:

2) I apologize for having lied to you. But I have absolutely no regret about having lied to you.

3) Thank you for the present. But I have no gratitude towards you for giving me the present.

One can imagine someone publicly uttering the first component of any of (1), (2), and (3) while uttering the second component *sotto voce*—in which case the public speech act (assertion, apology, and thanking, respectively) would simply be insincere. There is nothing very noteworthy about that. What is strange is when the second component is uttered *out loud* along with the first—since a speech act that wears its insincerity on its sleeve is apt to cause confusion.[13] It is, to quote Austin again, 'a statement that fails to get by' (1971, p. 18). Note that I am not merely saying that (1)-(3) 'sound odd to my ear,' but am trying to locate a particular kind of oddity, which promises to help us understand the relevant notion of *expressing* a mental state. David Copp introduces a useful phrase to denote this kind of expression relation: 'Moore-expression' (correspondence 2003; see his 2001, p. 10). It is illuminating, I think, to consider these matters when reflecting on pejorative language, such as racial slurs.

4) Aaron's a kike. But I have no contempt towards him or people of his ethnicity or religion.[14]

I am inclined to think that (4) is a manifestation of the same phenomenon as the rest, though Copp would disagree. Though he would concur that calling someone a 'kike' certainly in some sense expresses contempt, and that someone who uttered (4) should expect to be challenged to explain, he judges that it doesn't *Moore-express* contempt, since the implication can be canceled. (Copp's example concerns calling a canine a 'cur' as opposed to a 'mongrel dog'; I am assuming that it is a fair analogy.) One could add further explanatory comments to (4) so as to assuage the audience's confusion in the following manner: 'Aaron is a kike; but I have no contempt towards Aaron or people of his ethnicity or religion; it's just that for the moment I've forgotten the usual non-derogatory term for such persons.' Thus, Copp thinks, the contempt expressed by 'kike' may be canceled. Perhaps he

[13] Compare the phenomenon of sarcasm: If Fred looks out at the pelting English rain and sighs 'Another glorious day in paradise' in tones of sarcasm clear for all to hear, then he has not asserted that it is a glorious day in paradise. But sarcasm is such an entrenched convention that his audience will not be confused—they will know that he did not assert it, and did not intend to assert it, and Fred will know that they know. Thus sarcasm is not a case of insincerity 'worn on the sleeve,' since it is not a case of insincerity at all.

[14] A lot depends on tone of voice here. It does with (1)-(3) as well, since uttering any of these in tones of sarcasm or in a clearly joking way completely alters the speech acts involved. Let me stipulate that we are to imagine these uttered in serious tones. (Insert the word 'stinking' before 'kike' if it helps to reinforce this point.) I apologize to anyone who finds the term offensive even in the context of *mentioning the word as an example of an extremely offensive word.* On another occasion of making this argument, I employed the less provocative term 'kraut'—but the very choice of a word whose offensiveness is less deeply ingrained undermined the cogency of the argument.

is right, though I confess to feeling uneasy about having to rely on an example of someone forgetting a word or learning a language in order to illustrate the context of cancelation. Is there *anything* that is noncancelable if we admit such odd contexts?

Consider the case of the Reverend William Spooner, who famously concluded one of his sermons with the addendum: 'In the sermon I have just preached, whenever I said "Aristotle" I meant to say "St. Paul."' Suppose the sermon had included the assertion of the sentence 'Aristotle was born in Tarsus' (when in fact it was St. Paul who was born there), and, when queried on this by a bewildered student, Spooner had responded (coming momentarily to his senses, but oblivious of what he had just been saying) 'No, of course Aristotle wasn't born in Tarsus; he was born in Stagira!' Thus we have what appears to be a flat contradiction asserted—'Aristotle was born in Tarsus and it is not the case that Aristotle was born in Tarsus' (the first conjunct uttered as part of the confused sermon, and the second conjunct, being an answer to a question, not part of the sermon)—but one made intelligible by the addendum admitting a linguistic confusion. This possibility hardly undermines the fact that 'Aristotle was born in Tarsus' is semantically implied by 'Aristotle was born in Tarsus,' and that this implication is as noncancelable as they come! Recall that Paul Grice confessed to the unreliability of cancelability test when speakers are talking 'in a loose or relaxed way' (Grice, 1989, p. 44), and I suspect that imagining a person who has an imperfect grasp of the language, or who is forced to employ unconventional words because she cannot remember the correct ones, renders the test null.

We will return to the question of the status of (4) shortly, but for now apply these thoughts to a moral utterance:

5) Hitler was evil. But I do not believe that he was evil.

This seems a straightforward instance of Moore's Paradox, which strongly suggests that an ordinary freestanding utterance of 'Hitler was evil' is an assertion. If an utterance of the form 'But I don't have mental state M' can nullify the speech act that the preceding comment would otherwise perform, then it is natural to assume (ceteris paribus) that the preceding comment functions to express M.

So does noncognitivism stand refuted? Not by any means, for it seems that we can observe the same phenomenon, or at least a very similar one, if in (5) we substitute for 'belief' something conative. Which conative state we choose, and how we opt to describe it, may make a significant difference here, and there are various options. Copp (following Allan Gibbard) prefers to see the expressed conative state as *acceptance of, or subscription to, a moral standard* (Copp, 2001, p. 30). According to this stipulative use of 'subscription,' if a person actually does take this attitude, 'she is in a state of mind that, if effective, constrains and guides her planning so that she is motivated to some degree [to comply]' (Copp 2001: 30). Since no proposal more plausible suggests itself, let us use the following wording:

6) Hitler was evil. But I subscribe to no normative standard that condemns him or his actions.

I'm cautiously tempted to treat (6) the same as the others, to hold that all six of the numbered sentence pairs so far considered are Moore-paradoxical, meaning that in each the relation between the speech act of the first sentence and the mental state mentioned in the second sentence is one of Moore-expression. If this is true of (5) and (6), then the conclusion to be drawn is that an ordinary freestanding utterance of 'Hitler was evil' expresses *both a belief*

and a conative attitude. This would speak in support of a metaethical view that mixes aspects of traditional noncognitivism with components of traditional cognitivism—a view I favor.[15]

Copp, however, though agreeing that (1), (2), (3) and (5) reveal instances of Moore-expression, thinks that (6) should go with (4)—as *not* Moore-paradoxical—for the reason mentioned earlier: that the strangeness of (6) may be assuaged via additional commentary, and thus the customary relation between the utterance and the mental state in question may be canceled. He would find nothing unintelligible about an amoralist who says 'Hitler was evil; but I subscribe to no normative standard that condemns him or his actions. I just don't go in for morality; I believe in it all right, but I think it's a manipulative cultural invention that is best avoided.' I have two responses to Copp's claim here. The first is to deny that the amoralist's declaration is intelligible *in the relevant way*, and thus maintain grounds for holding that (6) is Moore-paradoxical. The second is to accept that (6) is not Moore-paradoxical, but explore whether there is another kind of *expression* involved—a kind that, while not being Moore-expression, is nevertheless robust enough to underwrite the noncognitivist element of moral judgments. Either avenue is sufficiently promising that the prospects for the mixed metaethical view to which I have just adverted seem encouraging. Discussion of the former response calls for a section to itself, after which I will turn to the latter response.

IV. Amoralist cancelation

There is, of course, a sense in which all of the numbered sentence pairs (1)-(6) are intelligible. They are grammatical and noncontradictory sentences. But so too is any Moore-paradoxical sentence pair (and many nonsensical sentences besides). Yet I maintain that they are unintelligible in the sense that someone hearing any such pair (someone who has not been primed in some special way, that is) would be unsure about what speech act has been performed by the first component.[16] Considering (6): Copp is correct that any perplexity that may be aroused may also be assuaged with further commentary—the kind of 'but-I-don't-go-in-for-morality' comment that was mentioned. We are familiar with this kind of amoralist; after all, didn't Plato's Thrasymachus proclaim something along similar lines?[17]

Yet care needs to be taken in our treatment of these pieces of additional commentary that render intelligible preceding speech that would otherwise be confusing. There's a sense in which *any* verbal nonsense can be rendered intelligible with the addition of '… and what I just uttered was a great example of verbal nonsense.' The crucial matter is whether the

---

[15] Such a mixed view has been maintained by C.L. Stevenson, R.M. Hare, and P.H. Nowell-Smith, among others. If one defines noncognitivism simply as the denial of cognitivism, then the two theories are, of course, contradictories. Similarly, if one defines cognitivism as the theory that moral judgments express *only* beliefs, then any 'mixed theory' will be excluded. However, if we think of expressivism as a positive proposal (about what moral judgments *do* express), and we drop the 'only' clause in both expressivism and cognitivism, then moral judgments may be *two* things: They may be assertions *and* ways of expressing conative attitudes.

[16] Cf John MacFarlane (2005, p. 334), who writes: 'Imagine someone saying: "I concede that what I asserted wasn't true, but I stand by what I said anyway." We would have a very difficult time taking such a person seriously as an asserter. If she continued to manifest this kind of indifference to established truth, we would stop regarding the noises coming out of her mouth as assertions.'

[17] Note how 'amoralism' is a term of art here. In the vernacular, 'amoralist' often denotes someone who rejects morality altogether, who doesn't believe in it at all. In recent philosophical debates, by contrast, it denotes someone who makes genuine moral judgments but lacks any motivation to comply (and the topic of the debate is whether the amoralist is even a possibility). I am using the term is a third way: to denote someone who tries to cancel the motivation-implicating aspect of a moral judgment. The debate here is not whether such agents exist, but whether they succeed in making moral judgments.

additional intelligibility-instating commentary leaves intact the apparent speech act performed by the first sentence. For example, suppose that A utters after a dinner party 'Well, *the plates* were nice'; we might presume that it is being conversationally implicated that the food was unpleasant. But this implicatum may be smoothly canceled, if A were to add: 'Of course, I don't mean to imply that the food was unpleasant; it was nice too.' Whatever wisps of strangeness might remain hanging over this pair of utterances will be dissipated if A explains: 'I was just so taken by the plates that for a moment I wasn't thinking about the food.' The important thing to notice is that at the end of all this explaining we are content that A's first sentence was indeed what we thought it was: an assertion that the plates were nice. By comparison, if B says 'The cat is on the mat, but I don't believe it,' then goes on to add '… and that's a good example of confusing language,' then although listeners may be comfortable with the total exchange, they will not know whether B has asserted that the cat is on the mat. The crucial question, then, is not whether the Thrasymachian amoralist who renders intelligible (6) by explaining that he 'just doesn't go in for morality' is someone we can make sense of (I concede that he is); the question is whether after the intelligibility-reinstating amoralist explanation the audience is confident that the speaker has made a genuine moral judgment. And about this I think there is substantial doubt.

Before discussing the amoralist further, it will be useful to remind ourselves how fluid and scrappy linguistic conventions can be. Let me draw attention to four general points.

First, a solid linguistic convention may be quite easily overridden by another. There is little doubt that the term 'slut' functions as a pejorative in English. Yet by introducing the overarching convention of *joking*—which may be achieved in a second by a shift in tone or a twitch of an eyebrow—one might in a playful manner say to a close female friend 'Oh, you're such a slut' with all offensiveness nullified. Yet even in these circumstances 'slut' continues to be a contempt-expressing term, for that, after all, is what makes the comment funny.

Second, many terms that function to express attitudes as well as beliefs will also have purely belief-expressing uses as well. The word 'queer' in the sense of *unusual and peculiar* remains neutral even if 'queer' in reference to homosexuality can be used as a term of derision. The word 'bastard' for a long while could be used descriptively to mean *illegitimate offspring*, even when it could also be employed as a term of insult. Similarly, most if not all of the terms centrally associated with moral judgment also can be used non-morally. Possibly the most fundamental term of moral appraisal is 'ought.' The 'ought' that appears in 'Mary morally ought to refrain from stealing' may express the speaker's subscription to norms that condemn stealing, but nobody is claiming this of the weather forecaster's utterance of 'It ought to rain tomorrow.' Similarly, to acknowledge that someone is 'a good assassin' is not to express any kind of endorsing attitude, whereas to claim that someone is 'a good person' is.

A third point to note about such conventions is that they can change very quickly. The terms 'idiot,' 'moron,' and 'cretin' were once respectable scientific labels; the term that largely replaced them—'mentally retarded'—is at present the subject of controversy. There will be transitional times when one should not claim with confidence either that there is or is not a linguistic convention according to which the term expresses an attitude.

A fourth point is that linguistic subgroups can create linguistic sub-conventions. Within certain gay circles, referring to one's gay friends as 'queer' may be neutral, though it may be highly insulting for an outsider to select that term. Much the same could be said about the use of the word 'nigger' among some African American subgroups. In some circles, to call

something ‘bad’ (in a certain tone of voice, perhaps) is a way of praising it, and in surfing lingo, to call a wave ‘wicked’ is to express admiration for its qualities. Such conventions may be sequentially embraced and overridden within a single conversation—or, indeed, within a single sentence: ‘Which of you bastards called this bastard a bastard?’[18]

What I hope these observations call attention to is the fact that although we might be able to *imagine* someone intelligibly advocating the amoralist line, it doesn’t show that there is not actually an entrenched convention according to which the use of moral terms expresses subscription to a norm. What we need to ask ourselves is whether any such imaginative act involves us thinking of aberrant subgroups, or people speaking in a joking, playful manner, or the speaker using something like a sarcastic or ironic tone of voice, or using a moral term in a non-moral manner, or introducing a new convention by example, or a world with slightly different linguistic conventions than we actually do have, or so on. When Milton’s Satan says ‘Evil, be thou my good’ the natural reading is that Satan is doing something tricky with language. A careful analysis of his comment would take too long here; it’s enough to note that although we know exactly what Satan is trying to communicate, we also recognize that the surface construction is paradoxical (and this, of course, is what gives the line its poetic power). Thus, that Satan’s comment should be intelligible doesn’t reveal that there is not actually a linguistic convention according to which to call something ‘evil’ is to express one’s subscription to a standard that condemns it. It is exactly this convention that Milton has exploited in a clever and mischievous way.

I believe the same thing can be said quite generally of the amoralist’s apparent cancelation of the conative component of a moral judgment: ‘Hitler was evil; but I subscribe to no normative standard that condemns him or his actions. I just don’t go in for morality; I believe in it all right, but I think it’s a manipulative cultural invention that is best avoided.’ Although a person would be *intelligible* if she said this, it seems to me that we would be left in serious doubt as to whether she has really *judged Hitler to be evil*. And the reason for this indecision, I think, is precisely that the careful explanation that the speaker offers of her position reveals that a degree of stipulative usage is being introduced. She is explicitly suspending a convention that is in place in regular language. But the fact that one can do this, and do it with ease, hardly shows that there is not actually such a convention, any more than the fact that I can say ‘For the next few minutes I will use the word ‘cat’ to stand for dogs’ (or ‘Cats, be thou my dogs’) reveals that there is some doubt concerning whether in English ‘cat’ denotes cats.[19]

---

[18] This sentence was reportedly uttered by the Australian cricket captain during the ‘bodyline series’ of 1932-3. The English captain came to the Australian dressing room to complain about one of his players having been called a ‘bastard’ during play. Bill Woodfull, the Aussie captain, turned to his team and uttered the memorable line. (I owe this example to the late David Lewis [correspondence 2000].)

[19] There is a well-known interpretation of the amoralist from R.M. Hare (1952, pp. 124-26, 167 ff.), according to which the amoralist’s statement is not literally a moral judgment at all, but rather is best read as having quotation marks round the term ‘morally ought.’ My view is not unlike this, though it is important to bear in mind that Hare’s amoralist utters something like ‘For me to steal would be wrong’ while having no motivation to refrain from stealing, while my amoralist says ‘For me to steal would be wrong, though I subscribe to no normative framework that condemns stealing.’ Given the careful verbal qualification that the latter offers, it seems to me quite plausible that something rather like quotation marks are being imposed. Copp objects that Hare’s view fails to accommodate the possibility of moralists and amoralists entering into moral debate (Copp, 2001, p. 13). If the amoralist says ‘Liberalism is a great evil’ and the moralist responds ‘No, liberalism is morally defensible,’ but in fact the former statement is equivalent to something like ‘It is considered hereabouts that liberalism is a great evil,’ then there is no real disagreement. But I don’t find this the reductio that Copp seems to think it to be. Perhaps any intuition we have that there can be genuine moral debate with amoralists just stems from the fact that the way they speak (making the quotation marks tacit) is apt to encourage us to

V. Frege-expression

I have just argued in favor of treating the relation between moral judgment and certain conative states as an instance of Moore-expression. The issue of where the line should be drawn between Moore-paradoxical and non-Moore-paradoxical utterances is difficult to settle, since the rules for how we should restrict the contexts in which cancelation may or may not be possible are undecided. But suppose that Copp is correct that (4) and (6) are not examples of Moore-paradox. Does it follow that there is no expressivist component to pejorative slurs and moral judgments? Copp doesn't think so, and nor do I.

Drawing inspiration from Frege's views on 'coloring,' Copp claims that pejorative terms and moral terms *Frege-express* mental states. Frege wrote that two words might have the very same sense and reference, and yet one might lend the utterance a 'coloring' (*Färbung*) that the other does not, such that choosing to use one word rather than the other (for example, 'kike' rather than 'Jewish person') might be 'unsuitable, as if a song with a sad subject were to be sung in a lively fashion' ([1892] 1997, p. 167). Frege's own example involves the word 'cur': he writes that 'whilst the word "dog" is neutral as between having pleasant or unpleasant associations, the word "cur" certainly has unpleasant rather than pleasant associations and puts us in mind of a dog with a somewhat unkempt appearance' ([1897] 1997, pp. 240-241]. When such coloring becomes an entrenched custom in the linguistic community—as is the case with words like 'kike' or 'slut'—then we can, according to Copp, consider the relation between the utterance and the proposition that the speaker has the attitude in question (for example, contempt) to be a variety of conventional implicature. Note, though, that Copp thinks that this expressiveness is cancelable, and in this he diverges from Grice, for whom, apparently, noncancelability is a feature by which conventional implicatures are to be distinguished from conversational implicatures. Though cancelable, these 'colorings' are a type of conventional rather than conversational implicature (for Copp) because they are detachable[20] and because to employ a colored term while lacking the attitude in question would be a *misuse* of the term. At the risk of sounding evasive, I prefer to sidestep the Gricean framework, if only because it strikes me as sufficiently unclear and contested that one only courts controversy in trying to apply it to new domains. I have already expressed my misgivings about the cancelability of colorings, but even if I agreed with Copp on this point, it seems to me imprudent to employ the term 'conventional implicature'—a term of art partially defined by reference to *noncancelability*—to categorize the phenomenon.[21]

Nevertheless, I am strongly inclined to agree with Copp (and Frege) on the general point that colorings are, or at least can be, a matter of *linguistic convention*. This may be a vague claim, but it is good enough for my present purposes. The contemptuous attitude of someone who uses 'kike' rather than 'Jewish person' is not merely an expectation that interlocutors

---

forget that they are not really making moral judgments at all. If I were to hear someone claim that liberalism is a great evil then I would want to protest; but if I were then to discover that this person had earlier asserted 'Evil, be thou my good,' then I should become quite confused as to what she thought about liberalism, and thus not at all confident that I should disagree.

[20] For Grice, the implicatum *p* is detachable from an utterance iff there are ways of saying the same thing that do not implicate *p* (see Grice 1989, p. 39). What Copp has in mind in saying that colorings are detachable is that instead of 'Aaron is a kike' one could say 'Aaron is Jewish,' and the latter, though saying the same thing (having the same sense and reference?), lacks the implicatum that the speaker has contempt.

[21] Copp himself admits that 'nothing [in my argument] turns on whether coloring is an example of conventional implicature or simply a phenomenon that is similar to conventional implicature' (2001, p. 23).

will have formed on the basis of past observation. In teaching the word 'kike' to a novice language-user, it would be intolerably negligent to refrain from mentioning the term's evaluative baggage. Indeed, someone who didn't know that 'kike' was a contempt-expressing term could legitimately be said not to understand the term properly at all, even if able competently to apply it to all and only Jewish people. Any such ignorant person would not require any tutoring concerning what it takes to be Jewish, but is in need of *linguistic* instruction.

Although the comparison threatens to be misunderstood if taken too far, I agree with Copp that moral language is in important respects like pejorative language. More precisely: the way that moral judgments express conative attitudes is very similar to, if not the same as, the way that pejorative terms express attitudes. The manner in which (6) fails to get by may or may not be precisely the same as the manner in which (5) fails to get by, but it is close enough as to make no difference to the general conclusion that I am trying to reach: that moral judgments express, as a matter of entrenched linguistic convention, both beliefs *and* conative attitudes. The traditional debate between the cognitivist and the noncognitivist has not taken into account such nuances as the distinction between Moore-expression and Frege-expression—indeed, has been scandalously casual about what it means to say that such-and-such judgments *express* mental state so-and-so—and thus, were we ultimately to conclude that moral judgments Moore-express beliefs and Frege-express attitudes, this could not be construed as a victory for either party. My main point is that a modicum of reflection on the issue reveals the traditional metaethical debate between the cognitivist and the noncognitivist to rest on a false dichotomy.

VI. Hume: expressivist, cognitivist, *and* skeptic?

Earlier I claimed that there is very little evidence that Hume advocated expressivism. This was not entirely true; the point I was trying to press is that the places where Hume has traditionally been read as promoting expressivism (or noncognitivism more generally) should not be construed that way. For the real hints of expressivism in Hume, one must look to where he discusses evaluative *language*. As a preliminary, we should remind ourselves that Hume did have at his disposal a remarkably forward-looking account of how indicative language can be used in non-assertoric ways. His sophisticated discussion of promising foreshadows J.L. Austin's…

> … there is *a certain form of words* … by which we bind ourselves to the performance of any action. This form of words constitutes what we call a *promise*. … When a man says *he promises any thing*, he in effect expresses a *resolution* of performing it. (*T* 3.2.5.10/522)

Regarding evaluative language, Hume is (occasionally) clear that there are entrenched conventions that associate conative states with certain words:

> Every tongue possesses one set of words which are taken in a good sense, and another in the opposite. (*EPM* 1.10/ 174)

> … when [someone] bestows on any man the epithets of *vicious* or *odious* or *depraved*, he … expresses sentiments, in which, he expects, all his audience are to concur with him. (*EPM* 9.6/272)

> … there are certain terms in every language which import blame, and others praise; and all men who use the same tongue must agree in their application of them. … This great unanimity is usually ascribed to the influence of plain reason, which … maintains similar sentiments in all men … But we must also allow, that some part of the seeming harmony in morals may be accounted for from the very nature of language. The word *virtue*, with its equivalent in every tongue, implies praise, as that of *vice* does blame; and no one, without the most obvious and grossest impropriety, could affix reproach to a term, which in general acceptation is understood in a good sense; or bestow applause, where the idiom requires disapprobation. (*Of the Standard of Taste* [1757] 1996: 134-5)

It is in such passages—few and far between as they are—that we find Hume the expressivist. But these comments are presented in a way that makes clear that Hume considers them as peripheral to any of his central arguments; he evidently does not, on these occasions, take himself to be putting forward any weighty and controversial metaethical thesis in need of argumentative support. In other words, to the extent that Hume is an expressivist, it is not something he thinks worthy of making a song and dance about; he hardly notices that he is taking (what we would now classify as) a metaethical stance. More importantly, note that these expressivist musings have no obvious role to play in the moral thesis that Hume is really obsessed with: that morals are the product a sentimental faculty rather than a rational faculty (a thesis I earlier called 'psychological emotivism'). In particular, these passages are very far indeed—both logically and textually—from the much-touted *motivation argument* for expressivism.

What I hope to have shown in the previous two sections is that even if Hume does have expressivist leanings, this does not exclude his also robustly endorsing a cognitivist metaethical view. Moral cognitivism comes in both realist and skeptic flavors, and each of these possibilities is compatible with expressivism. Copp, for example, articulates and advocates a position he calls 'realist-expressivism.' He thinks that the truth conditions for the belief expressed by a moral judgment like 'Cursing is wrong' concern cursing being prohibited by a 'relevantly justified or authoritative moral standard or norm' (2001, p. 27). What it takes for a moral standard to be appropriately 'authoritative' is, in the first instance, left open. Copp's own view is a *society-centered theory*: that a standard is authoritative just in case 'its currency in the social code of the relevant society would best contribute to the society's ability to meet its needs—including its needs for physical continuity, internal harmony and cooperative interaction, and peaceful and cooperative relations with its neighbors' (2001, p. 28). Since, we may assume, such justification is sometimes forthcoming, on Copp's view moral judgments will turn out sometimes to be true. He combines this realism with the thesis that moral language Frege-expresses conative states; hence: realist-expressivism.[22]

I should like to draw attention to another branch of the tree of metaethics: *error theoretic expressivism*. There are different ways that one might argue for this position, but it is convenient to use Copp's view as a point of departure. Suppose he is correct that the cognitive element of a moral claim refers to something like the 'relevantly justified or authoritative moral standard or norm.' One may, nevertheless, think that Copp's preferred explication of *justification* is too relativistic or too anthropocentric to capture the kind of practical authority we demand of a moral theory. For the kind of familiar reasons outlined by John Mackie, for example, one might think that inherent in moral discourse is a commitment

[22] I actually harbor some reservations that Copp's view quite deserves the label 'realism,' but the fact that it is a version of moral cognitivist 'success theory' is enough to underwrite the distinction I am highlighting. For my views on how to characterize *moral (anti)realism*, see Joyce, 2007*a*.

to a kind of institution-transcendent practical categoricity that is in fact not satisfied by anything in the world (Mackie, 1977; see also Joyce, 2001).[23] Thus one might agree with Copp concerning how to understand the expressivist element of moral discourse, and also agree with his views concerning the truth conditions (*broadly* construed) of moral judgments, while holding that these truth conditions are never satisfied. Hence: error theoretic expressivism.

Might Hume be a realist-expressivist or an error theoretic expressivist? I wouldn't want to press either case with any confidence, but I will nevertheless close with a brief exploration of Hume's commitment to the moral error theory. Note that I do not claim for a moment that Hume thought of himself as an error theorist but just expressed himself poorly, nor even that he would have embraced the view had it been articulated to him. But there are certainly threads in Hume's moral philosophy that lean in that direction.[24]

In looking for evidence *against* this interpretation, one might bring forth any of a number of Hume's critical comments aimed at moral skepticism. But it should be remembered that the kind of skeptic whom Hume has in mind is invariably the *Pyrrhonic* skeptic: someone who thinks that we cannot know whether claims of a certain kind are true or false and therefore ought to withhold passing judgment on the matter. The error theorist, by contrast, is no Pyrrhonic skeptic, but (in classical terms) should be classified as a negative dogmatist (or nihilist). Similarly, many of Hume's comments apparently targeting moral nihilists (for example, 'those who have denied the reality of moral distinctions' [*EPM* 1.2/169]) on more careful examination seem to be admonishing those who would pretend *indifference*, who would claim not to *care* whether a person was honest or a thief. But this is also something that a moral error theorist may distance himself from. The moral error theorist may be as opposed to tax fraud, as sickened by pedophilia, as horrified by genocide, as anyone else. Error theoretic moral skepticism implies nothing about how *tolerant* its advocates will be.

Hume is opposed not merely to Pyrrhonic ataraxic indifference towards morality, he is averse to any suggestion that philosophizing should lead us to *give up* the practice of making moral judgments. I can discern no hint of moral eliminativism in his writings. The widespread assumption that eliminativism is the natural consequence of a moral error theory may have something to do with a reluctance to press the error theoretic interpretation of Hume. This assumption, however, is flawed. There may be pragmatic reasons for maintaining moral thought and moral language even once moral skepticism has been embraced (see Joyce, 2001; Kalderon, 2005). Or it may be that the human mind is simply *unable* to give up these practices, even when philosophical considerations have led one to see the flaws. One can find allusions to the former in Hume's writings, and the latter is something of a recurring theme. In his essay 'The Sceptic,' he notes 'that famous doctrine' that colors exist not in nature but only in the eye, then poses the rhetorical question: '[If this were so,] would dyers or painters ever be less regarded or esteemed?' He goes on to ask 'why should a like discovery in moral philosophy make any alteration?' ([1742] 1996: 354).

---

[23] Indeed, Mackie's general definition of 'good' is not a million miles away from the cognitivist element of morality articulated by Copp. Mackie defines 'good' as 'such as to satisfy requirements (etc.) of the kind in question' (1977, pp. 55-6). With many non-moral uses of 'good' Mackie thinks the predicate is satisfied. But in *moral* contexts, he thinks, the pertinent requirements are those that are 'simply there, in the nature of things, without being the requirements of any person or body of persons, even God' (p. 59). It is Mackie's conviction that there are no such 'intrinsic requirements' that leads to his moral skepticism.

[24] Mackie (1980) argues along similar lines to me for the error theoretic interpretation of Hume. David Gauthier (1992) also toys with this interpretation of Hume's account of the artificial virtues, though doesn't firmly endorse it.

The closing section of Book 1 of the *Treatise* waxes lyrical about how simply *living* life will drive all skeptical musings from one's mind:

> I dine, I play a game of back-gammon, I converse, and am merry with my friends; and when after three or four hour's amusement, I wou'd return to these speculations, they appear so cold, and strain'd, and ridiculous, that I cannot find it in my heart to enter into them any farther. (*T* 1.4.7.9/269)

In his abstract for the *Treatise* (written in the third person singular), Hume sums up this view:

> Our author insists upon several other sceptical topics; and on the whole concludes, that we assent to our faculties, and employ our reason only because we cannot help it. Philosophy would render us entirely *Pyrrhonian* were not nature too strong for it. (*T* 657)

Of anyone who would profess indifference to moral distinctions, Hume councils that the best response is 'to leave him to himself,' trusting that 'it is probable he will, at last, of himself, from mere weariness, come over to the side of common sense and reason' (*EPM* 1.2/170). In other words, even if one came to espouse a moral error theory, 'nature herself' would eventually drive that philosophical allegiance from one's mind. And Hume evidently thinks that this would be no bad thing. Of those 'honest gentlemen of England'—who 'being always employ'd in domestic affairs, or amusing themselves in common recreations, have carried their thoughts very little beyond those objects'—Hume tells us that 'they do well to keep themselves in their present situation' (*T* 1.4.7.14/272). Yet even if these down-to-earth folk are acknowledged to be well off, Hume is not attempting to dissuade anyone from engaging in philosophical speculations: He simply thinks that one either will or will not, according to temperament and mood, and that to the extent that one will it is likely that one's efforts will be temporary (that is, until someone calls out 'Anyone for backgammon?').

Yet none of this is at odds with the possible *truth* of an error theory. Regarding causal relations and the continued existence of external objects, Hume is explicit that experience 'leads us into errors' (*T* 1.4.7.6/267) (for he thinks that the two beliefs are jointly affirmed contraries), but even here—where we seem to have Hume clearly endorsing some kind of error theory—he thinks it remains an open question the extent to which he should 'torture my brain with subtilities' (*T* 1.4.7.10/270), the extent to which he should 'yield to these illusions' (*T* 1.4.7.6/267). Nature may ineluctably reassert herself against the awareness that one has fallen into error, but the errors are no less errors for that.

Might things stand similarly for morality, in Hume's eyes? He certainly never claims outright that morality doesn't exist. He is more likely to say something along the lines of 'Moral properties exist not in bodies but merely in the senses.' It is worth noting that such claims remain consistent with a moral error theory. Telling someone 'The pink elephants exist only in your mind' is in fact a way of saying that the pink elephants do not exist at all. At one point in the *Treatise* Hume declares that sounds and smells 'really exist no where' (*T* 1.3.14. 20/167), and it is reasonable to think that he will say the same of color and causation (the section in question concerns 'necessary connexion'). This bald claim of non-existence comes immediately after he has spoken of the mind's 'great propensity to spread itself on external objects'—a thesis that I will here refer to as 'projectivism.' The other well-known projectivist passage from Hume is in *EPM*, where he claims 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' (*EPM*

appendix 1.21/294). Here he mentions 'beauty and deformity, vice and virtue' as the products of sentimental projection. The correlation of these two projectivist passages suggests that what goes for one (that is, that sounds and smells 'really exist no where') should go for the other (that is, that virtue and vice 'really exist no where'). That moral qualities should receive the same treatment as color, sound, smell, heat and cold is reaffirmed elsewhere in the *Treatise* (3.1.1.21/469) and in 'The Sceptic.'

Hume is, moreover, explicit that the folk do indeed think of colors, smells, sounds, and heat as objective qualities of objects. In a letter to Hugh Blair of 1762 he is dismissive of the idea that the folk might not be objectivists; evidently, 'the Vulgar' (as Hume refers to them) are taken in by their own projectivist tendencies. 'Philosophy scarce ever advances a greater Paradox in the Eyes of the People, than when it affirms that Snow is neither cold nor white: Fire hot nor red' ([1762] 1986, p. 416). On the assumption that what is said here will carry over for moral qualities as well, then the folk are generally fooled by their moral projectivist tendencies: They are unaware that their moral judgments are the product of sentiments being projected onto the world; they both experience the world as morally 'colored' *and believe it to be*. But if someone *believes* something to be the case, then it is natural to assume that her utterances on the matter will be *assertions*. Thus, I claim, Hume's moral projectivism is a form of psychological emotivism that leads naturally to (but I would not go so far as to say *implies*) moral cognitivism—though a cognitivism that remains compatible with an expressivist component.[25]

But are the assertions in question *true*? If snow is not white—something that Hume seemingly endorses—and someone asserts the sentence 'Snow is white,' then the very natural conclusion to draw is that she has simply asserted something false. Similarly, if Fred's character does not have the quality of *virtuousness*, and someone asserts the sentence 'Fred is virtuous,' then the natural conclusion is that she has asserted something false. One way to avoid this error theoretic interpretation of Hume's metaethical commitments is to problematize the analogy that allows us to draw conclusions about his implicit moral views from what he somewhat more explicitly claims about sensory modalities. I have nothing to say here on that score, except to reaffirm that Hume draws the analogy sufficiently frequently that it is safe to assume that he thinks there are illuminating similarities among these topics. Another way is to leave the analogy intact but to deny the error theoretic construal of both analogs. Perhaps when Hume denies that snow is white what he means is something like 'Snow—*considered in itself, restricting ourselves to the consideration of only its intrinsic qualities*—is not white.' But perhaps he also thinks that it is not obligatory to understand the truth conditions of 'Snow is white' in this manner. If we allow the possibility of whiteness being some relational, subject-implicating property, then perhaps Hume will consider the sentence true, after all. The problem is that he has adamantly affirmed that the general folk think of the whiteness of snow in the former objectivist fashion. The crux of the issue, then, is whether the weight of dominant folk belief on the matter is sufficient to determine how the word 'white' must be understood (or, if you prefer, how the identity conditions of the concept *whiteness* must be construed). Understood one way it leads to an error theory, understood in another way it leads to a success theory.

Hume had no resources for addressing this question, and nor have we. This leads to an impasse in establishing whether Hume's occasional expressivist tendencies are mixed with a cognitivism that is committed to success or a cognitivism that is committed to skepticism.

---

[25] In saying this I am flying in the face of recent tradition, which tends to lump projectivism together with noncognitivism. I have argued elsewhere that this is at best optional and at worst a mistake. See Joyce, 2006 (ch.4) and Joyce 2009. In the latter, I delineate different species of projectivism.

The contemporary debate between the moral error theorist and the moral success theorist seems locked up at exactly the same point: A problematic (or 'queer') quality of morality is brought forth, regarding which some will argue that this quality is an *essential* aspect of the moral conceptual framework (such that any normative system stripped of this problematic element would no longer deserve the name 'morality'), whereas others will argue that extirpation of the flawed element would amount merely to a benign revision and demystification of morality. Lacking an accepted methodology for deciding such disputes, the modern metaethical debate is at a disappointing stalemate.[26] It is also possible that on occasions there is no fact of the matter about whether a given discourse (for example, morality) is committed to some putatively queer property, leading to the conclusion that the dispute between the moral error theorist and her many detractors may in fact be fundamentally undecidable.[27] If this is correct, then it is tempting to suppose that any interpretation of Hume that leaves him sitting on the fence over this matter is a charitable one.[28]

[26] I discuss this impasse in Joyce, 2006 (ch.6), 2007*a*, and 2007*b*.

[27] David Lewis writes: "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, p. 93).

[28] Early portions of this chapter closely follow Joyce, 2002. A youthful version of this paper once went by the name of 'Noncognitivism, Motivation, and Assertion,' and it was helped along by feedback from David Lewis and Simon Kirchin. I thank David Copp for very useful discussion.

Bibliography

Árdal, P.S. (1966) *Passion and Value in Hume's Treatise* (Edinburgh, University of Edinburgh Press).

Austin, J.L. (1962) *How to Do Things with Words* (Oxford, Oxford University Press).

Austin, J.L. (1971) 'Performative-Constative,' in J. Searle (ed.) *The Philosophy of Language* (Oxford, Oxford University Press), pp. 13-22

Austin, J.L. [1970] (1990) 'Performative Utterances,' in A.P. Martinich (ed.) *The Philosophy of Language* (New York, Oxford University Press), pp. 105-114.

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

Ayer, A.J. (1980) *Hume* (New York, Hill and Wang).

Blackburn, S. (1993) *Essays in Quasi-Realism* (Oxford, Oxford University Press).

Bricke, J. (1996) *Mind and Morality: An Examination of Hume's Moral Psychology* (Oxford, Oxford University Press).

Carnap, R. (1935) *Philosophy and Logical Syntax* (London, Kegan Paul, Trench, Trubner & Co. Ltd.).

Cohen, R. (1997) 'Is Hume a Noncognitivist in the Motivation Argument?' *Philosophical Studies* 85, 251-266.

Copp, D. (2001) 'Realist-Expressivism,' *Social Philosophy and Policy* 18, 1-43.

Fessler, D., Arguello, A., Mekdara, J. & Macias, R. (2003) 'Disgust Sensitivity and Meat Consumption: A Test of an Emotivist Account of Moral Vegetarianism,' *Appetite* 41, 31-41.

Flew, A. (1963) 'On the Interpretation of Hume,' *Philosophy* 38, 178-181.

Frege, G. [1892] (1997) 'On Sense and Reference,' in M. Beaney (ed.), *The Frege Reader* (Oxford, Basil Blackwell), pp. 151-171.

Frege, G. [1897] (1997) 'Logic,' in M. Beaney (ed.), *The Frege Reader* (Oxford, Basil Blackwell), pp. 227-250.

Gauthier, D. (1992) 'Artificial Virtues and the Sensible Knave,' *Hume Studies* 18, 401-427.

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.

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.

Grice, P. (1989) *Studies in the Way of Words* (Cambridge, MA, Harvard University Press).

Haidt, J. (2001) 'The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment,' *Psychological Review* 108, 814-834.

Hare, R.M. (1952) *The Language of Morals* (Oxford, Oxford University Press).

Hare, R.M. (1999) *Objective Prescriptions and Other Essays* (Oxford, Oxford University Press).

Harman, G. & Thomson, J.J. (1996) *Moral Relativism and Moral Objectivity* (Oxford, Blackwell).

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

Hume, D. [1742] 1996. 'The Sceptic.' In *David Hume: Selected Essays* (Oxford, Oxford University Press), pp. 95-113.

Hume, D. [1751] 1998. *An Enquiry Concerning the Principles of Morals*. Oxford, Clarendon.

Hume, D. [1757] 1996. 'Of the Standard of Taste.' In *David Hume: Selected Essays* (Oxford, Oxford University Press), pp. 133-154.

Hume, D. [1762] (1986) 'A New Letter to Hugh Blair from July 1762,' *Mind* 95, 411-16.

Joyce, R. (2001) *The Myth of Morality* (Cambridge, Cambridge University Press).

Joyce, R. (2002) 'Expressivism and Motivation Internalism,' *Analysis* 62, 336-344.

Joyce, R. (2006) *The Evolution of Morality* (Cambridge, MA, MIT Press).

Joyce R. (2007*a*) 'Moral Anti-realism,' *The Stanford Encyclopedia of Philosophy*.

Joyce R. (2007*b*) 'Morality, Schmorality,' in P. Bloomfield (ed.) *Morality and Self-Interest* (Oxford, Oxford University Press), pp. 51-75.

Joyce, R. (2008) 'What Neuroscience can (and cannot) Contribute to Metaethics,' in W. Sinnott-Armstrong (ed.) *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (Cambridge, MA, MIT Press), pp. 371-394.

Joyce, R. (2009) 'Is Moral Projectivism Empirically Tractable?' *Ethical Theory and Moral Practice* 12: 53-75.

Kalderon, M. (2005) *Moral Fictionalism* (Oxford, Oxford University Press).

Lewis, D.K. [1989] (2000) "Dispositional Theories of Value," in his *Papers in Ethics and Social Philosophy* (Cambridge: Cambridge University Press), pp. 68-94.

MacFarlane, J. (2005) 'Making Sense of Relative Truth,' *Proceedings of the Aristotelian Society* 105, 321-339.

Mackie, J. (1977) *Ethics: Inventing Right and Wrong* (New York, Penguin Books).

Mackie, J. (1980) *Hume's Moral Theory* (New York, Routledge).

Malcolm, N. (1958) *Ludwig Wittgenstein: A Memoir* (Oxford, Oxford University Press).

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.

Moore, G.E. (1942) 'A Reply to my Critics,' in P.A. Schilpp (ed.) *The Philosophy of G.E. Moore* (Evanston, IL, Northwestern University), pp. 535-677.

Price, H. (1988) *Facts and the Function of Truth* (Oxford, Blackwell).

Prinz, J. (2007) *The Emotional Construction of Morals* (Oxford, Oxford University Press).

Ridge, M. (2006) 'Sincerity and Expressivism,' *Philosophical Studies* 131, 487-510.

Searle, J.R. (1969) *Speech Acts* (Cambridge, Cambridge University Press).

Shafer-Landau, R. (2005) *Moral Realism* (Oxford, Oxford University Press).

Smith, M. (1994) *The Moral Problem* (Oxford, Oxford University Press).

Snare, F. (1975) 'The Argument from Motivation,' *Mind* 84, 1-9.

Snare, F. (1991) *Morals, Motivation, and Convention* (Cambridge, Cambridge University Press).

Stevenson, C.L. (1937) 'The Emotive Meaning of Ethical Terms,' *Mind* 46, 14-31.

Stevenson, C.L. (1963) *Facts and Values* (New Haven, CT, Yale University Press).

Sturgeon, N. (2008) 'Hume's Metaethics: Is Hume a Moral Noncognitivist?' in E.S. Radcliffe (ed.) *A Companion to Hume* (Oxford, Blackwell).

Svavarsdóttir, S. (2006) 'How do Moral Judgments Motivate?' in J. Dreier (ed.) *Contemporary Debates in Moral Theory* (Oxford, Blackwell), pp. 163-181.

Timmons, M. (1999) *Morality Without Foundations* (Oxford, Oxford University Press).