

# Are ethical judgments intrinsically motivational? Lessons from “acquired sociopathy” [1]

ADINA ROSKIES

---

**ABSTRACT** *Metaethical questions are typically held to be a priori, and therefore impervious to empirical evidence. Here I examine the metaethical claim that motive-internalism about belief (or belief-internalism), the position that moral beliefs are intrinsically motivating, is true. I argue that belief-internalists are faced with a dilemma. Either their formulation of internalism is so weak that it fails to be philosophically interesting, or it is a substantive claim but can be shown to be empirically false. I then provide evidence for the falsity of substantive belief-internalism. I describe a group of brain-damaged patients who sustain impairment in their moral sensibility: although they have normal moral beliefs and make moral judgments, they are not inclined to act in accordance with those beliefs and judgments. Thus, I argue that they are walking counterexamples to the substantive internalist claim. In addition to constraining our conception of moral reasoning, this argument stands as an example of how empirical evidence can be relevantly brought to bear on a philosophical question typically viewed to be a priori.*

## 1. Introduction

Empirical evidence is generally thought to be irrelevant to philosophical theorizing in a number of philosophical areas, including metaethics. One metaethical issue often held to be immune from the empirical concerns the relation between moral facts and moral motivation. I argue here that the motive-internalist about belief (or belief-internalist)—one who claims that moral beliefs entail motivation—faces a dilemma. Either his internalist thesis is a weak one and as such does not interestingly characterize the domain of the moral, or his thesis is a philosophically interesting one, but demonstrably false. In order to argue for its falsity, I describe neuropsychological data from patients with injury to ventromedial (VM) cortex and argue that VM patients are walking counterexamples to this internalist thesis. This analysis suggests that motive-internalism about moral belief is not a tenable position. Moreover, it suggests that moral philosophy need not be, and perhaps ought not be, exclusively *a priori*.

*Adina Roskies, Department of Philosophy, Massachusetts Institute of Technology, E-39, 55 Hayward Street, Cambridge, MA 02138, USA, email: adinad@mit.edu*

## 2. Internalism

Motive-internalism, generally conceived, holds that motivation is intrinsic to, or a necessary component of, moral belief or judgment. According to the internalist view, moral belief or judgment alone contains or entails moral motivation: recognition that one ought, morally, to *A* constitutes motivation to *A* [2]. So, for instance, the belief that I ought to give money to famine relief guarantees that I am, to some degree, motivated to do so. In contrast to the internalist, the externalist maintains that moral belief or judgment can exist independently of any motivational force. To be sure, to prompt action, that moral belief or judgment must be accompanied by a corresponding moral desire, but this desire is one which is only contingently related to the moral belief.

Three characteristics of the internalist claim are worthy of note: necessity, intrinsicness, and specificity. First, necessity: internalism says not merely that it happens to be the case that motivation accompanies belief or judgment, but rather that it must be so. Thus, internalist claims are modal claims; they purport to state a necessary truth about ethics. Internalism should therefore not be a thesis that pertains solely to humans, but rather to any agent capable of moral beliefs. Second, internalism has a characteristic I call intrinsicness: internalist claims are committed to the view that the connection between moral belief and motivation must hold in virtue of the content of the moral belief itself, not in virtue of some contingent or auxiliary non-moral fact or reason. Nagel (1970, p. 7) describes this criterion in *The possibility of altruism*: “Internalism is the view that the presence of a motivation for acting morally is guaranteed by the truth of ethical propositions themselves. On this view the motivation must be so tied to the truth, or meaning, of ethical statements that when in a particular case someone is (or perhaps merely believes that he is) morally required to do something, it follows that he has a motivation for doing it” (see also Mele, 1996; Smith, 1993, Chapter 3). Mele (1996, p. 730) uses the term “motivation constituting” for this feature of moral beliefs. The motivation to act in accordance with one’s beliefs or judgments, in other words, must stem from the moral character of a belief or judgment itself [3]. Third, internalist positions were originally inspired by the conviction that moral facts are different in kind from other facts, or that moral reasoning differs in specific ways from non-moral reasoning. Internalism was originally intended to explain or describe these differences. It is still the case that interesting forms of internalism attempt to set moral beliefs off from other varieties of belief: the motivational force of moral belief distinguishes them as special, since other types of beliefs are not intrinsically motivating (Mele, 1996, p. 732). Thus, the internalist will maintain that anyone who sincerely believes that morality dictates that he or she ought to give money to famine relief must thereby be motivated to give, although no such related motivation must attend the sincere belief that the law dictates that he or she ought to pay his or her taxes. This feature, which I call specificity, is not diagnostic of internalism, but it is central to the spirit of internalism, and characterizes the most philosophically interesting versions of it. In what follows, I will consider a claim relating moral belief to motivation to be an internalist claim only if it possesses the first two characteristics, necessity and

intrinsicness; I further maintain that the interest of any such claim diminishes considerably if the third criterion, specificity, is not also fulfilled.

In this paper I argue that the motive-internalist about belief faces a dilemma: either internalism is a rather weak and philosophically uninteresting thesis about the nature of moral belief, or else internalism is false. Before proceeding further, I lay the groundwork with a few preliminaries:

1. For the purposes of this paper, I will assume moral cognitivism to be true: moral sentences are truth-apt; moral beliefs are a species of belief [4].
2. Internalist theses are sometimes couched in terms of belief, and other times in terms of judgment. Since judgment plausibly entails belief, I take it that if I provide arguments sufficient to refute belief-internalism, these also, *a fortiori* refute judgment-internalism.
3. When I talk about moral belief, I will assume that the belief is a first-person “ought” belief. That is, I recognize that the internalist is not committed to saying that if I hold a moral belief that someone else ought to do *A*, this entails that I am motivated to do *A*. To be fair to the internalist, we will assume that internalist hypotheticals are first-person claims.
4. The most common forms of internalism are belief- or judgment-internalism. Some Kantians, however, argue that reasons are intrinsically motivating. Here I will refrain from considering reason-internalism. Whether a rationalist internalist thesis is tenable remains an open question.

### 3. The dilemma

A variety of different theses have been offered in the spirit of motive-internalism. Some formulations of internalism I take to be philosophically uninteresting, at least insofar as they pertain to ethics, for either they fail to be internalist claims at all, or their interest is diluted because they apply equally to many non-ethical realms as well. Among these are formulations that tie internalism to practical rationality, and ones that rely upon claims of normalcy.

The former internalist thesis holds that there is a necessary connection between judgment and motivation, except in cases in which the agent’s practical rationality is compromised:

PI: If an agent believes that it is right to  $\Phi$  in circumstances *C*, then either he is motivated to  $\Phi$  in *C* or he is practically irrational. (Smith, 1993, p. 61)

While we may ultimately want to accede to something like PI, on the face of it PI as an internalist thesis is too weak to be revealing about the nature of moral judgment. Briefly, PI is unsatisfactory without a further account of what it is to be practically rational. For if, as is often held, to be practically rational is merely to desire to act in accordance with what one judges right or best, then PI is trivially true. It can always be satisfied, regardless of evidence or argument, provided that one is always ready to conclude that an agent is practically irrational. However, notice that so construed PI does not seem to be a strong claim about the nature of

morality, but rather a definitional claim about the rational status of agents. Furthermore, PI lacks specificity: it is a claim that applies to non-moral construals of what is best as well as to moral ones. I may well think it best to set the table with my fork on the left; if I am not motivated to do so I am, according to PI, practically irrational. So if PI is to be a claim about the nature of moral reasoning, the defender of PI must say something more than that practical irrationality is failure to act in accordance with one's judgments of what is right or best to do.

A second type of claim that I find philosophically anemic goes something like this:

UN<sub>1</sub>: Usually/normally, if an agent believes that it is right to  $\Phi$  in circumstances  $C$ , then he is motivated to  $\Phi$  in  $C$ .

Or like this:

UN<sub>2</sub>: If an agent believes that it is right to  $\Phi$  in circumstances  $C$ , then usually he is motivated to  $\Phi$  in  $C$ .

Now the problem with this type of claim is that it loses its philosophical force. With the addition of the qualifying "usually" or "normally," the internalist thesis becomes a sneaky way of saying PI, or it becomes a statistical claim. As statistical claims, the UN theses are rather odd. Is internalism true if moral motivation accompanies desire in 51% of the cases, but false if it only happens 49% of the time? It is incumbent upon the proponent of this formulation of internalism to say what "usually" or "normally" means. On some readings of the UN theses, they threaten to fail to meet the necessity criterion. If the thesis need only hold in the actual world, then according to UN<sub>1</sub> or UN<sub>2</sub>, in another possible world a being could hold moral beliefs yet not be motivated to act in accordance with them—the truth of internalism would vary by world. Or, suppose we take the UN claims to be defining of morality. Consider two similar worlds which differ in the extent to which some people are motivated to act in accord with their beliefs. Will we be forced to say of intrinsic duplicates in these worlds that one's belief is moral, while the other's psychophysically identical belief is not, solely because of the frequency with which others are motivated by their beliefs? This threatens to undermine the relation of motivation to the content of the belief. Finally, one worries that the "usually" phrase turns the internalist claim from a metaphysical one about an essential aspect of moral judgment to a merely descriptive claim about what generally is the case.

One might hold an internalist view akin to UN<sub>2</sub>, with the additional rationale for thinking it true that our moral practices only exist as they do because we are generally motivated to act in accordance with our moral beliefs [5]. Thus, the existence of our practices of using moral language, making moral judgments, etc. ensures the truth of the internalist connection between morals and motivation. However, this quasi-transcendental argument for internalism falls short of the criteria for internalism that I have laid out. For even if our practices do depend upon something like UN<sub>2</sub> being true, such a claim plausibly fails to meet the intrinsicness criterion—the claim is consistent with a situation in which the moral beliefs themselves are not the motivating factor. For example, it is plausible to maintain that our

practice of placing colored lights at busy intersections depends upon the fact that we are motivated to stop at red lights. Indeed, we would not use traffic lights were we not generally inclined to modify our driving habits to comply with their conventional meaning. So, in this sense, our traffic light usage depends upon the connection between the presence of traffic lights and our motivations, just as our use of moral language may depend upon the fact that we are often inclined to modify our behavior to accord with moral beliefs. Nonetheless, it is implausible to hold that this truth about the practice of setting traffic signals gives us reason to believe that something about the nature of the lights, their redness perhaps, intrinsically motivates stopping behavior, or is intrinsically motivational at all. Rather, it is far more plausible to hold that we are motivated to stop because of factors entirely extrinsic to the nature of the signal. Analogously, even if our use of moral language depends upon a reliable connection between moral belief and action, we still lack reason to believe that this is because of a special feature associated with the moral nature of the belief. The burden of proof is thus placed upon the proponent of internalism, who must say more to convince us that his position meets our criteria [6].

None of these versions of internalism, then, makes the substantive claim about the nature of moral judgment that internalism, in its original formulations, makes. PI lacks specificity, and the UN claims, if they meet the criterion of necessity at all, plausibly fail to meet the intrinsicness criterion. ~~Moral functionalism, although it provides an account of how it is that motivation often accompanies moral belief, poses problems for the internalist and offers at best an account of why some moral beliefs are sometimes motivating.~~ Thus, I claim that interesting forms of internalism must avoid being merely descriptive claims about what we usually do, or definitional claims about practical reason. Let us call an internalist thesis that does this a substantive internalist thesis. An example of the substantive internalist thesis, then, is this:

SI: If an agent believes that it is right to  $\Phi$  in circumstances  $C$ , then he is motivated to  $\Phi$  in  $C$ .

I will argue now that this substantive internalist claim is empirically false. To do so, I will introduce a group of humans with damage to the ventromedial frontal area of the brain, who I take to be walking counterexamples to the strong internalist claim.

#### 4. The neuropsychology of VM damage

Ventromedial (VM) prefrontal cortex is anatomically connected to a wide variety of brain areas, including those associated with perception, reasoning, declarative knowledge, and with emotion and visceral control. VM cortex is therefore uniquely functionally situated to mediate between the neural systems for arousal and emotion, and those supporting linguistic cognition.

The first reported case of VM damage was in 1848. Phineas Gage was a man who seemed headed for certain success. He had a promising career as a railway construction foreman, was well respected by peers and superiors alike, and enjoyed a stable and happy family life. But, in one of the most bizarre accidents in the annals

of medicine, a premature detonation of explosives sent a spear-like tamping iron rocketing through Gage's skull and out the top of his cranium (Damasio, 1995; Damasio *et al.*, 1994). Amazingly, Gage survived his injury, and moreover, suffered no impairment of his intellectual functions. But although his memory, reasoning, speech and motor functions remained intact, something was severely affected—so much so that his doctor reported, “Gage was no longer Gage” (Damasio *et al.*, 1994). After his injury, Gage made poor choices, acted inappropriately in public, behaved irresponsibly, and could not hold a job. He appeared to have a deficit that selectively affected his behavior in social situations. He failed to act reliably in socially acceptable ways, regardless of whether those were delineated either by widely held ethical judgments or by social convention.

Phineas Gage is perhaps the most dramatic example of a class of brain-damaged individuals who have sustained damage to a common part of their neural architecture, the ventromedial part of prefrontal cortex [7]. Like Gage, VM patients appear cognitively normal on a wide spectrum of standard psychological tests, including those measuring intelligence and reasoning abilities, and indeed, in casual encounters these patients may seem normal. However, VM patients all appear to have particular difficulty in acting in accordance with social mores, despite their retained ability to judge appropriately in such situations. The functional dissociations which VM patients exhibit present a unique opportunity to probe the functional relationships between social and moral reasoning and action. A group of VM patients have been studied extensively by neurologist Antonio Damasio and colleagues (see Adolphs *et al.*, 1996; Bechara *et al.*, 1997, 2000; Saver & Damasio, 1991). Damasio's case report of patient EVR further illustrates the disorder:

By age 35, in 1975, EVR was a successful professional, happily married, and the father of two. He led an impeccable social life, and was a role model to younger siblings. In that year, an orbitofrontal meningioma was diagnosed and, in order to achieve its successful surgical resection a bilateral excision of orbital and lower mesial cortices was necessary. EVR's basic intelligence and standard memory were not compromised by the ablation. His performances on standardized IQ and memory tests are uniformly in the superior range [97–99th percentile]. He passes all formal neuropsychological probes ...

Standing in sharp contrast to this pattern of neuropsychological test performance, EVR's social conduct was profoundly affected by his brain injury. Over a brief period of time, he entered disastrous business ventures (one of which led to predictable bankruptcy), and was divorced twice (the second marriage, which was to a prostitute, only lasted 6 months). He has been unable to hold any paying job since the time of the surgery, and his plans for future activity are defective. (Damasio *et al.*, 1990)

Damasio and colleagues refer to this condition as “acquired sociopathy.” They have developed a battery of laboratory tests to further characterize the condition. Here I summarize the results of several studies.

1. VM patients are able to make appropriate moral and social judgments when

queried. When presented with hypothetical situations, the conclusions they reach about moral questions concur with those which normals typically reach. Psychological evaluation shows that some VM subjects attain the highest level of abstract moral reasoning [8].

2. Clinical histories and observation suggest that VM patients are impaired in their ability to act effectively in many moral situations.
3. Normal subjects produce a skin-conductance response (SCR) to emotionally-charged or value-laden stimuli [9]. In contrast, VM patients do not generally produce SCRs when presented with such stimuli [10]. However, other tests produce normal SCRs in VM patients, demonstrating that the autonomic nervous system itself is undamaged.
4. VM patients display and report attenuated or absence of affect when faced with situations that reliably elicit emotions in normals.

For the purposes of this paper, I take a measurable SCR to be evidence of the presence of motivation, and lack thereof to be indicative of absence of motivation [11]. This simplification is warranted: the presence of the SCR is reliably correlated with cases in which action is consistent with judgment, and its absence is correlated with occasions in which the VM patient fails to act in accord with his or her judgments. Thus, the SCR is a reliable indicator of motivation for action [12]. I conclude this section with a summary of the three primary features of note in this clinical profile:

- **KNOW:** VM patients retain the declarative knowledge related to moral issues, and appear to be able to reason morally at a normal level. Significantly, their moral claims accord with those of normals.
- **ACT:** VM patients fail to reliably act as normals do in many ethically charged situations.
- **MOTIVATE:** in ethically-charged situations, VM patients seem to lack appropriate motivational and emotional responses, both at the level of their own subjective experience, and with regard to normal physiological correlates of emotion and motivation.

## 5. Do VM patients have a moral deficit?

Before using VM patients as a test case for issues in ethics, we must establish whether the social dysfunction of the VM individuals is, in fact, moral. After all, they are variously described as having difficulty with decision making, with choosing correctly in social and moral contexts, and with moral conduct.

There are several objections that come to mind in light of the complex nature of the VM neuropsychological profiles. For example, one may question whether lack of moral motivation on the part of VM patients merely reflects a general impairment in motivation, or listlessness (see Mele, 1996; Stocker, 1979). However, there is considerable specificity in the motivational defects of these patients. VM patients retain appetitive motivations such as the motivation to seek food and eat when hungry, they appear to want the company of others, and they seem to want to

succeed financially, as is evidenced by their attempts to win money in a gambling task and their (usually ill-fated) financial schemes. Thus, while certain kinds of motivation are impaired, others, often those with explicit and immediate rewards, remain intact. Ethical motivations are a subset of the impaired motivations.

Another possibility is that the deficit these subjects have is not in complying with moral judgments *per se*, but in acting in relation to evaluative judgments more generally. However, this too is rather unlikely. Behavior in relation to aesthetic judgments, and judgments of gustatory taste, for example, appear to be unaffected [13]. There is evidence that their impairment involves action as it relates to the expectation of delayed reward or punishment. This may indicate a deficit in prudential motivation, which is often involved in motivation in ethical situations.

Moreover, to maintain that the deficits of VM patients are not moral, one would have to appeal to a very weak view of ethics, in which only violations of the most central and indisputable ethical tenets constitute moral failings. A stronger view of ethics is more reasonable, where a myriad of more subtle actions, such as keeping promises, discharging one's responsibilities and telling the truth, are also counted as moral actions. VM patients fail regularly in these lesser, quotidian moral demands. In fact, it is because of these failures that VM patients are sometimes described as "hurting themselves more than others," for these breaches of trust erode the relationships that form the foundation of the social structures by which we so often measure a person's successes or failures. Given the strong view of ethics, the VM patients' ethical behavior is indeed differentially impaired, even though it typically manifests itself in a socially innocuous way.

Finally, to expect a person with a moral deficit to be an amoral monster may be unrealistic. Is it not more likely that a person unmotivated to act morally will, in normal circumstances, fail to act morally in small ways? After all, in everyday life the moral decisions that we are called upon to make are typically minor ones. To be a real moral monster would require very *strong* motivation to act *immorally*, and it is precisely this type of socially-conceived goal-directed motivation that seems to be impaired in VM patients.

Nonetheless, one may still be puzzled by the rarity of violent acts committed by VM patients, and the suspicion may arise that moral motivation prevents them from being truly immoral. However, another explanation for the VM behavior is available: their general lack of violence may be due to behavioral habits acquired prior to their injury, which, in a morally blind way, prevent them from the most egregious infringements of the moral code. In support of this, VM brain-damaged subjects who acquired their injury early in life are more prone to violence, and inflict harm upon others without signs of empathy or remorse (Anderson *et al.*, 1999). These patients with early damage fail to acquire the declarative knowledge of social and ethical norms which enable them to judge morally or act morally; they never acquire any behavioral habits associated with moral judgment. The neural and motivational defects appear to be comparable in the cases of early and late damage; if cases of early VM damage are found to be morally relevant, so should the cases of later damage.

## 6. Substantive belief-internalism

Now, let us reconsider SI, the substantive internalist thesis, in light of the VM data:

SI: If an agent believes that it is right to  $\Phi$  in circumstances  $C$ , then he is motivated to  $\Phi$  in  $C$ .

VM patients appear to have moral beliefs: they sincerely assert moral sentences that are in accord with our own; they have moral knowledge and thus moral belief. However, VM patients do not reliably display the signs of motivation in situations in which moral beliefs are affirmed. Thus, VM patients have moral beliefs, but lack the motivation which normally co-occurs with such beliefs. Taken together, KNOW + MOTIVATE show that VM patients are counterexamples to the claim that there is a necessary connection between moral beliefs and motivation.

Note that this analysis goes well beyond the rather obvious fact that moral beliefs don't always lead to action. Failure to act is suggestive of, but not proof of, lack of motivation. This analysis can discriminate between cases where agents fail to act because of lack of motivation, or because of countervailing motivations. Recall that SCR is present in normals in cases associated with motivation. The SCR response is present in the situation regardless of whether the agent acts in such a situation or not. If the SCR is a sign of motivation, and such motivation were to underlie moral behavior associated with avoiding such circumstances, then we would expect to see signs of it regardless of the presence of other motivations which prevent action. Indeed, one might expect that conflicting motivations, though they may prevent action, would result in a higher magnitude SCR than an unopposed motivation since both negative and positive valenced motivation lead to SCRs. However, since in VM patients the SCR is notably absent, we can conclude that both moral motivation and conflicting motivation are absent.

This simple argument is sufficient to refute other varieties of belief-internalism. For example, one formulation of belief-internalism holds that the intrinsic connection between moral belief and motivation is one of identity: desires just are a species of belief, and moral beliefs are instances of such "besires." Lewis has argued that the "desire as belief" thesis conflicts with the principles of decision theory and is therefore false (Lewis, 1988, 1996). VM data provide an independent empirical argument against the thesis. For if moral beliefs just are motivational states, motivation must necessarily co-occur with moral beliefs. The dissociation between moral belief and moral motivation in VM patients demonstrates that the thesis must be empirically false.

## 7. The internalist's objections

The committed internalist will argue that although the VM patients have beliefs and make judgments that *seem* moral, they are not *really* moral, and thus VM patients are irrelevant to internalist claims. Let us examine whether such an objection can be upheld.

This objection revolves around the issue of whether VM patients know the

meaning of moral terms. I will argue that it is not plausible to hold that VM patients lack mastery of moral language. Before their brain damage, VM subjects clearly had mastery of moral terms, for they were just like you or me. Is it to be thought that mastery, a form of knowledge, is undone by their type of brain damage? According to all tests, their language and declarative knowledge structures are intact. VM patients exhibit no cognitive deficits, either with memory or language. There is no evidence, then, that VM damage disrupts any form of knowledge, and thus, if they initially knew the meaning of moral terms, they still do. Analogously, although it might be maintained that congenitally blind people never achieve full mastery of color terms, it is far harder to argue that a newly blind person lacks mastery of the term “red.” At the extreme, one would have to argue that blindfolded people, or people with their eyes closed lose mastery of color terms. Thus, a plausible argument that VM patients lack mastery of moral terms will have to include an account of how and why their previous mastery of moral language deteriorates.

Furthermore, there is no suggestion that the moral concepts of VM patients change as a result of their brain damage. We have already seen that VM patients’ judgments appear to be the same as those of normals. *Prima facie*, if the initial premises and the reasoning processes which underlie their considered beliefs are those which a competent moral agent would acknowledge as moral, then one would presume the beliefs to be moral. To argue that VM patients’ beliefs and judgments are not really moral, despite the fact that they appear to correspond with normals’ moral beliefs and judgments, would be to argue that they have different content. However, this is not plausible. Given the above argument and the fact that VM patients are fully normal in their moral reasoning (remember, some VM patients reach the highest developmental stage of moral reasoning), it is difficult to see how the contents of their beliefs would differ from normals. Indeed, the fact that normal cognitive operations result in outputs apparently identical to those normals produce strongly suggests that the objects over which the operations range are the same as in normals.

It is implausible, then, to think that the deficit exhibited by the VM patients is a failure in knowledge. Still, the skeptic may argue that the VM patients only have moral beliefs and judgments in the “inverted comma” sense. That is, he may hold that the VM patient is merely expressing what he takes other people’s moral beliefs or judgments to be (Hare, 1956, p. 124). However, VM patients claim they are relating their own beliefs or judgments, and, as noted above, there is no evidence that the VM patients’ grasp of meanings has changed. It would be particularly difficult to argue that they now fail to understand the meaning of the question, “What do you think?” and instead take it to mean, “What do others say?” Moreover, there is no indication of any duplicity on the part of the VM patients. There is no evidence or cause for insincerity in their behavior, they respond to experimenters similarly about their moral and non-moral beliefs, and they consider themselves to be as morally aware and upstanding as anyone else. Unlike with the amoralist, there is no *prima facie* reason to expect them to deceive.

I think these considerations show the objector’s position to be highly implausible. Thus, I maintain that SI is empirically false.

## 8. Prospects for internalism

These arguments do not rule out all forms of belief-internalism, but they significantly constrain the space of possibilities. Here I will consider three other possible formulations of belief-internalism that might be thought to withstand the evidence presented above. The first I argue fails to qualify as an internalist thesis at all, though as a statement it might be true. The second suggests an interesting framework within which to consider the relation of moral belief to motivation, but I present reasons to doubt whether the connection it posits between these qualifies as a robust version of internalism. The final one is similar to the second in spirit; I will mention it only as a possible direction in which an internalist can proceed, since an account for motive-internalism has not yet been worked out in the literature.

Some may argue that the evidence I have thus far presented speaks for a usually/normally version of internalism, but one that wasn't addressed earlier. VM patients are brain-damaged, they will argue—and that they are an anomaly speaks to the truth of a version of internalism I will call UN<sub>3</sub>:

UN<sub>3</sub>: If a normal agent believes it is right to  $\Phi$  in circumstances  $C$ , then he is motivated to  $\Phi$  in  $C$ .

If UN<sub>3</sub> were true, it would be a scientifically interesting and significant statement about moral cognition [14]. However, as a philosophical internalist thesis, UN<sub>3</sub> seems to be lacking. One problem is that a normal agent is hard to define without resorting to statistical means, or without some understanding of the biological function of ethical reasoning. We must take care not to beg the question and define normalcy to be “motivated to act according to one’s moral judgment.” Furthermore, on the face of it, UN<sub>3</sub> seems to fail to satisfy either the necessity or the intrinsicness criterion. The internalist must justify that a claim like UN<sub>3</sub> is not merely a descriptive claim. While UN<sub>3</sub> might well be true of us (now), what reason do we have to suppose that this must be the case? Unless the proponent of UN<sub>3</sub> can present an argument for necessity, UN<sub>3</sub> does not stand as a robust internalist thesis. The same can be said for the intrinsicness criterion: the internalist must go further and demonstrate that UN<sub>3</sub> is true in virtue of ethics. Nagel (1970, p. 7) reminds us that “Externalism is compatible with the view that such motivation is always present—so long as its presence is not guaranteed by moral judgments themselves, but by something external to ethics.” I contend that the data from VM patients can establish, *prima facie*, that the content of moral judgment is independent of the neural structures subserving motivation, and thus that moral belief and motivation are dissociable. The burden of proof rests therefore on the internalist, who must demonstrate in what way motivation is intrinsic to ethics itself, and not just to the extrinsic emotional system that happens, in us, to be contingently connected to the cognitive moral system.

A second theory, moral functionalism, has been offered to explain the apparent connection between moral belief and motivation (Jackson & Pettit, 1995). Moral functionalists claim to occupy a middle ground between internalism and its denial, providing an explanation for why moral beliefs often appear to be motivating, but

allowing that they need not always be. Moral functionalists hold that moral terms have their meaning in virtue of their role in a network of mutually interdependent terms. Although atomistic definitions of such terms will not be forthcoming, we can evaluate whether an agent has mastery of networked terms by whether or not he subscribes to certain platitudes associated with the role of those terms. So, for example, someone who understands the word “fair” will assent to platitudes such as “If one alternative is fair, and if other things are equal, then that is the right option for the agent to desire and pursue,” along with a host of others (Jackson & Pettit, 1995, p. 23). Jackson and Pettit (1995, p. 24) argue that believing propositions containing moral terms “means becoming disposed to draw the lessons articulated in the commonplaces.” Insofar as these commonplaces involve desire or dispositions to act, motivation is built in to the meaning of moral terms.

So far moral functionalism seems to be merely an explication of internalism. However, Jackson and Pettit hold that the internalist connection is defeasible. Moral functionalists argue that evaluative content supervenes on descriptive content, and propose that there are two ways of understanding the same functionally specified content: one can understand such meanings intellectually, by knowing how they are situated in the network of terms which gives individual terms their meanings; and one can understand them non-intellectually, by being disposed to infer, judge, or act in a way consistent with that specified by appropriate platitudes [15].

Importantly, understanding in these two ways can come apart: people can understand moral terms intellectually, but not non-intellectually, and vice versa. Jackson and Pettit argue that the non-intellectual way of understanding the moral is canonical, and intellectual understanding is parasitic upon it. They could therefore explain the VM data as follows: VM patients lack the non-intellectualist mode of understanding moral terms, and are therefore not motivated by their moral beliefs, yet because of their preserved intellectual mode of understanding, they nonetheless have mastery of moral terms. According to moral functionalism, moral beliefs may or may not be necessarily motivating, depending upon the way in which moral content is grasped.

Here I wish only to point to moral functionalism as an internalist position potentially consistent with the VM data. There are many questions such a position raises, not the least of which is whether moral functionalism is the right characterization of the meaning of moral terms, which is ultimately an empirical question. In addition, Jackson and Pettit acknowledge that some platitudes may have to be abandoned in the ongoing process of building coherent cognitive networks [16]. Most internalists contend that one of the *a priori* commonplaces about moral terms is the statement of internalism itself. But, if Jackson and Pettit are right, we may find that this commonplace, if such it is, is one we must reject. Some empirical studies already indicate that normal people are not inclined to judge moral terms according to the internalist platitude (Nichols, 2002). If the internalist platitude is rejected, internalism for the moral functionalist can only hold on a piecemeal basis: only beliefs in those moral propositions containing moral terms associated with unrejectable platitudes that are connected to action could be held to be intrinsically motivational. This falls short of the spirit of internalism, which purports to make a

claim about moral beliefs generally. There are, finally, reasons to doubt whether such a position qualifies as an example of the robust internalism against which this paper is directed. Depending upon how one cashes out the theory, it seems to fail the intrinsicness and/or specificity criteria [17].

A third approach that could reconcile internalism with empirical data showing that moral beliefs are not necessarily motivating would have to argue that internalism is a necessary yet defeasible thesis. Lance and Little (in preparation) articulate a view that necessity and defeasibility are not mutually exclusive, and that necessary generalizations that admit ineliminably of exceptions can be genuinely explanatory and can illuminate the nature of phenomena by privileging certain of their aspects. According to their view, the motivational power of moral belief would be privileged, and it is by this privileging that we can view cases in which moral beliefs are not motivational as anomalies. This view is not unlike moral functionalism, in that the connection between moral belief and motivation is taken as privileged or canonical. A thorough exploration of this is beyond the scope of this paper; but it remains to be seen whether a *necessary* generalization that admits irreducibly of exception makes sense, and whether there is independent motivation to privilege those cases of moral belief that are accompanied by motivation.

## 9. Conclusion

I have argued that the motive-internalist about belief is faced with a dilemma: either the thesis is weak and fails to say something substantive about moral reasoning, or the thesis is false. The practical reason version of internalism is really a claim about practical reason, not moral judgment, and “normally” claims are merely descriptive. The substantive internalist thesis I have argued is falsified by VM patients, who demonstrate that having moral beliefs is not sufficient for motivation. In support of this, I have argued that the VM deficit is relevantly moral, and that it is not plausible to argue that VM patients don’t really have moral beliefs. The data from these patients allow me to argue that motivation, not merely action, is indeed absent in these patients, thus going beyond the type of argument that akrasia can present for internalism. Furthermore, I argue that these patients are not subject to the objections from internalists that revolve around the issue of whether they have mastery of moral concepts. The VM patients therefore go beyond other counterexamples to internalism in that they are real, not hypothetical; they demonstrably lack motivation, not merely sufficient motivation to act; and there are defensible arguments to the effect that they have mastery of moral terms. This analysis importantly constrains what we take to be the nature of moral beliefs, as well as the space of possibilities for other formulations of internalism. To the extent that examination of this type of data can help us constrain the space of possible internalist positions, this paper illustrates the important methodological point that empirical data can be relevant to questions in metaethics.

## Acknowledgements

I have benefited greatly from discussions with Sarah Buss, Tyler Doggett, Elizabeth Harman, Bernhard Nickel, Philip Pettit, Georges Rey, Judith Thomson, and Ralph Wedgwood. This work was supported by a grant from the McDonnell Project in Neurophilosophy, funded by the James S. McDonnell Foundation.

## Notes

- [1] This paper was the winner of the 2002 William James Prize of the Society for Philosophy and Psychology, awarded for the best paper by a graduate student presented at the Society's annual meeting.
- [2] See Mele (1996) and Smith (1993, 1995) for a discussion of different varieties of internalism. I will be concerned with what Mele calls "robust internalism" about ethics.
- [3] Mele (1996, p. 732) describes these two criteria as follows: "Believing oneself to be morally required to *A* metaphysically or conceptually (hence internally, in a formal sense) guarantees that one has a motivation for *A*ing; second, what is guaranteed, more precisely, is that motivation for *A*ing is built into any belief that one is (oneself) morally required to *A* and is internal to belief of that kind in this sense."
- [4] Many recent contributions to the internalism debate have assumed moral cognitivism (for a discussion, see Darwall *et al.*, 1992). The arguments presented here do not suffice to refute internalism in a non-cognitivist framework.
- [5] Simon Blackburn argues for something like this, but within a non-cognitivist framework.
- [6] Mele (1996) discusses reasons for thinking that the link between moral belief and motivation might not be of the right sort for internalism, even if the two routinely co-occur; he points out that the burden of proof rests on the internalist, not his opponent.
- [7] Because of the critical involvement of ventromedial cortex in this syndrome, I will often refer to people with such damage as VM patients.
- [8] Administration of Kohlberg's moral reasoning tests revealed that EVR had attained a late conventional/early post-conventional stage of reasoning. For comparison, note that by age 36, only 11% of American males reach the stage of post-conventional moral reasoning (Saver & Damasio, 1991).
- [9] An SCR is a physiological measure of arousal, such as is used in lie detector tests. It is also often called the galvanic skin response (GSR).
- [10] EVR did not produce SCRs to target (emotionally-charged) stimuli that he viewed passively. If instructed to respond verbally to the stimuli, he produced a normal SCR (Damasio *et al.*, 1990, p. 88).
- [11] This is a simplification on a number of fronts. First, it assumes an interpretation of the physiological response that maps neatly onto our psychological terminology, when the reality is unlikely to be so neat. It is possible that a large class of actions, such as simple bodily movements, and which we philosophically take to be motivated, do not require involvement of hypothalamic systems. It is also unknown whether habitual complex actions require the proper functioning of the autonomic nervous system. Such questions can be resolved empirically. Second, the difficulties in measuring SCR leave it open that some autonomic response may be present yet sub-threshold. Nonetheless, I think such a simplification is admissible since comparisons between situations in which an SCR is present with similar situations in which it is absent correlate well with behavioral differences.
- [12] It is a matter of important future work to elucidate more clearly what motivation consists in. For present purposes, I take it that motivation is akin to a species of desire, not necessarily in the sense of intense yearning (my desire for a large portion of French fries), but in the sense sufficient to impel us to action. For instance, although I do not desire to pay my taxes in the same way that

I desire to eat a large portion of French fries, I nonetheless am moved to pay them. It is this attenuated form of desire that I intend when I speak of when I speak of motivation. The arousal signaled by the SCR is minimally necessary for this type of motivation.

- [13] It must be noted that such dimensions were not explicitly tested in controlled situations; a more focused study of the domain of their impairment may lead to further insight into the nature of their disorder.
- [14] But see, for instance, Stocker (1979) and Mele (1996) for arguments that the truth of such a statement doesn't necessarily imply internalism.
- [15] For example, believing modus ponens intellectually would consist in endorsing the principle that "q follows from *if p then q* and *p*." But one can also believe modus ponens in a non-intellectual mode, simply by being disposed to make inferences in accordance with the principle, even if one lacks the explicit representation of the principle itself.
- [16] "The commonplaces that emerge in this process as those whose rejection cannot be countenanced will be taken to fix the relevant roles; they are the *a priori* compulsory propositions that anyone who knows how to use the terms is in a position to recognize as true. Other commonplaces—other putatively *a priori* propositions—will have to be dismissed as false or downgraded to the status of empirical, contingent truths" (Jackson & Pettit, 1995, p. 26).
- [17] First, it is important to recognize that according to this account it is not moral beliefs or moral content, meaning or truths, per se, that is intrinsically motivating, but rather content together with the way of grasping it that is. Thus, what seems crucial to motivation is the mode of understanding, which is plausibly a factor extrinsic to morals: this threatens to fail to satisfy the intrinsicness criterion. Indeed, as Jackson and Pettit argue, these two ways of understanding can be said to apply to many of our beliefs, moral, logical, sensible, and so on. Furthermore, Jackson and Pettit concede that the two ways of understanding the same moral content involve the mastery of different concepts. So, for instance, they allow that someone may understand modus ponens without mastery of the concept of entailment. But if this is so, there are two sets of moral concepts, those that are involved in grasping moral content non-intellectually, and those that are involved in grasping it intellectually. Only concepts belonging to the former set are intrinsically motivational. It seems attractive to many to hold that moral content must be psychologically grasped to be motivating, so that it is the moral concepts, rather than content defined independently of concepts, that we ought to look to as a source of motivation. This poses the following problem for the moral functionalist *qua* internalist: if he insists it is content, not concepts that matter, he has to qualify his internalism to allow that only moral beliefs held in a certain way are motivating; he has to explain how content, rather than concepts are the relevant factor; and he has to justify in what sense the descriptive content that our motivational concepts pick out is the moral content that our platitudes make reference to. If, on the other hand, the moral functionalist embraces the notion that it is grasp of moral concepts that is intrinsically motivating, other problems arise. Since only a subset of moral concepts are motivational, this view seems to fail to satisfy the specificity desideratum, and it also leads us to question whether the intrinsicness criterion is satisfied—in virtue of why is this set and not the other motivational, if not because of its moral nature? On the other hand, the moral functionalist *qua* internalist might restrict what he considers "moral" to the motivational subset. This whittling down of the moral may conflict with our common sense understanding. Here too, we need an account of why the refined domain of the moral is motivational, for it is plausible that motivation stems from some extrinsic factor, instead of the content of the moral belief. This issue promises to be hard to address from a cognitivist perspective.

## References

- ADOLPHS, R., TRANEL, D., BECHARA, A., DAMASIO, H. & DAMASIO, A.R. (1996). Neuropsychological approaches to reasoning and decision-making. In A.R. DAMASIO, Y. CHRISTEN & H. DAMASIO (Eds) *Neurobiology of decision-making* (pp. 157–179). Heidelberg: Springer-Verlag.
- ANDERSON, S.W., BECHARA, A., DAMASIO, H., TRANEL, D. & DAMASIO, A.R. (1999). Impairment of

- social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, 2, 1032–1037.
- BECHARA, A., DAMASIO, H., TRANEL, D. & DAMASIO, A.R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275, 1293–1295.
- BECHARA, A., DAMASIO, H. & DAMASIO, A.R. (2000). Emotion, decision-making and the orbitofrontal cortex. *Cerebral Cortex*, 10, 295–307.
- DAMASIO, A.R. (1995). *Descartes' error: emotion, reason and the human brain*. New York: Avon Books.
- DAMASIO, A.R., TRANEL, D. & DAMASIO, H. (1990). Individuals with sociopathic behavior caused by frontal damage fail to respond autonomically to social stimuli. *Behavioral Brain Research*, 41, 81–94.
- DAMASIO, H., GRABOWSKI, T., FRANK, R., GALABURDA, A.M. & DAMASIO, A.R. (1994). The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science*, 264, 1102–1105.
- DARWALL, S., GIBBARD, A. & RAILTON, P. (1992). Toward fin de siècle ethics: some trends. *Philosophical Review*, 101, 115–189.
- HARE, R.M. (1956). *The language of morals*. Oxford: Clarendon Press.
- JACKSON, F. & PETTIT, P. (1995). Moral functionalism and moral motivation. *Philosophical Quarterly*, 45, 451–472.
- LANCE, M. & LITTLE, M. (in preparation). Mad dogs and Englishmen: moral valence, defeasibility, and privileged conditions.
- LEWIS, D. (1988). Desire as belief. *Mind*, 97, 323–332.
- LEWIS, D. (1996). Desire as belief II. *Mind*, 105, 303–313.
- MELE, A. (1996). Internalist moral cognitivism and listlessness. *Ethics*, 106, 727–753.
- NAGEL, T. (1970). *The possibility of altruism*. Oxford: Clarendon Press.
- NICHOLS, S. (2002). How psychopaths threaten moral rationalism: is it irrational to be amoral? *Monist*, 85, 285–304.
- SAVER, J.L. & DAMASIO, A.R. (1991). Preserved access and processing of social knowledge in a patient with acquired sociopathy due to ventromedial frontal damage. *Neuropsychologia*, 29, 1241–1249.
- SMITH, M. (1993). *The moral problem*. Oxford: Blackwell.
- SMITH, M. (1995). Internalism's wheel. *Ratio*, 8, 277–302.
- STOCKER, M. (1979). Desiring the bad: an essay in moral psychology. *Journal of Philosophy*, 76, 738–753.