Daniel Eggers

Unconditional motivational internalism and Hume's lesson

1. Introduction

Motivational internalism is known as the view that there is an internal or necessary connection between moral judgement and motivation. In contrast, motivational externalists claim that any connection between moral judgement and motivation is external and contingent, moral motivation being dependent, for instance, on a prior desire to do the (morally) right thing. The most simple and straightforward form of motivational internalism, which I will refer to as unconditional motivational internalism (see also Francén 2010; and Strandberg 2012), holds that a person who makes a moral judgement is motivated to act in accordance with that judgement, no matter whether or not any further conditions are satisfied:

> *Unconditional motivational internalism* (UMI): necessarily, if a person judges that it is morally wrong to ϕ, then she is, at least to some extent, motivated to refrain from ϕ-ing.

The current metaethical debate is dominated by weaker forms of motivational internalism. The reason for the predominance of these more complex internalist views is that UMI is widely considered to pose too strong a connection between moral judgement and motivation. Thus both motivational externalists and professed internalists such as Michael Smith and

Jamie Dreier criticize that the simple version of motivational internalism cannot account for well-known cases in which moral motivation is defeated – cases of weakness of will or depression, for example. Therefore, Smith, Dreier and others propose versions of motivational internalism in which the necessary connection between moral judgement and motivation is tied to further conditions, the most popular proposals referring to conditions of psychological normalcy, practical rationality and virtuousness:

> *Conditional motivational internalism* (CMI): necessarily, if a person judges that it is morally wrong to ϕ, *and* if she is psychologically normal/practically rational/virtuous, then she is, at least to some extent, motivated to refrain from ϕ-ing.

In the present paper, I want to argue that the widespread rejection of UMI in favour of CMI is premature because, in effect, UMI makes a much weaker claim than the available attempts to refute it suggest. The commitments of UMI, and in particular the implications of the restriction built into UMI's motivational claim, have hitherto not been explicated with the appropriate amount of detail. However, once we explicate what it means to be only *to some extent* motivated, UMI is not the obviously implausible claim that it is often taken to be but a claim that demands serious consideration.

In what follows, I will first try to spell out the true commitments of UMI. In order to do so, I will start from another famous debate in the history of moral psychology – the debate over psychological egoism and psychological altruism –, and from David Hume's critique of psychological egoism in particular. In order to show what UMI really amounts to, I will draw upon one of Hume's thought experiments, originally meant to illustrate the true (and in Hume's view quite modest) commitments of psychological altruism. Armed with a more precise understanding of UMI, I will then go on to ask whether empirical evidence can help us decide whether UMI is true or false.

2. Hume and psychological egoism

It seems helpful to turn to the egoism/altruism debate in discussing motivational internalism because psychological altruism has suffered a fate that is somewhat similar to the one suffered by UMI. Psychological altruism, i.e. the claim that human beings are sometimes motivated altruistically, has often been taken for the different, and stronger, claim that human beings sometimes *act* altruistically. In other words, psychological altruism has not only been taken to mean that there is such a thing as altruistic motivation, but to also imply that, at least sometimes, this altruistic motivation overrides all other motivations and issues in action. However, as is already pointed out by Hume, psychological egoism in its philosophically interesting form is already defeated once we allow altruistic or benevolent motivation in human beings at all. Accordingly, Hume's defence of altruism explicitly allows for the possibility that altruistic motivation may, after all, always be overridden by egoistic motives and hence never issue in altruistic action (see Hume 1777, 270–71; see also Hume 1777, 225–26).

Hume supports his critique of psychological egoism with an instructive thought experiment. We are asked to imagine a being with absolutely no concern for others that is to assess two possible states of the world: the "prosperity" and the "ruin" of nations. As Hume points out, this being would necessarily be indifferent with regard to the two options and thus unable to choose between them.

> Let us suppose a person originally framed so as to have no manner of concern for his fellow-creatures, but to regard the happiness and misery of all sensible beings with greater indifference than even two contiguous shades of the same colour. Let us suppose, if the prosperity of nations were laid on the one hand, and their ruin on the other, and he were

> desired to choose; that he would stand like the schoolman's ass, irresolute and undetermined, between equal motives; or rather, like the same ass between two pieces of wood or marble, without any inclination or propensity to either side. The consequence, I believe, must be allowed just, that such a person, being absolutely unconcerned, either for the public good of a community or the private utility of others, would look on every quality, however pernicious, or however beneficial, to society, or to its possessor, with the same indifference as on the most common and uninteresting subject. (Hume 1777, 235)

Yet, as Hume is keen to emphasize, if we imagine human beings as we know them in the same situation, we would expect them to have a clear preference for the prosperous state of affairs, even where they would not personally benefit from it.

> But if, instead of this fancied monster, we suppose a *man* to form a judgement or determination in the case, there is to him a plain foundation of preference, where everything else is equal; and however cool his choice may be, if his heart be selfish, or if the persons interested be remote from him; there must still be a choice or distinction between what is useful, and what is pernicious. (*Ibid.*)

According to Hume, this expectation is only intelligible once we allow the existence of non-egoistic motives: In expecting a human being to choose the prosperous state of affairs even where this yields no personal benefit to him or her, we commit ourselves to the view that unlike the "fancied monster" human beings are capable of altruistic or benevolent motivation – even if this motivation may only lead to altruistic or benevolent *actions* in highly artificial scenarios.

3. Hume's lesson

As with psychological altruism, one gets the feeling that what most critics of UMI dismiss, and some of them quite easily, is not UMI, but, in fact, the stronger thesis that somebody making a moral judgement will necessarily act in accordance with that judgement.[1] To be sure, virtually all opponents of UMI explicitly acknowledge that UMI *does not* make this stronger claim, but claims only *some* motivation on behalf of the person making the judgement. However, the actual treatment that most critics of UMI allow their target suggests that they do not remain faithful to their explicit definition of UMI or are at least not fully aware of how powerful the limitation built into UMI actually is. Thus the examples that are used as counterexamples to UMI are usually designed in ways that appear inappropriate or, at the very least, inconclusive once we take this limitation seriously.

In some cases, the complete absence of moral motivation required to refute UMI is simply stipulated without any further justification.[2] Case descriptions that thus take the absence of moral motivation for granted, however, will only seem plausible to someone who already thinks that externalism is true. They are hardly suited, therefore, to demonstrate the superiority of one of the two positions over the other. The same goes for rejections of UMI which do not refer to concrete cases at all, but only point out that UMI is incompatible with certain phenomena, such as weakness of will or depression. Since the exact nature of these phenomena is contentious, reference to them does not provide much help in trying to establish the inappropriateness of UMI, either.[3]

[1] For a similar criticism, see Garrard and McNaughton 1998, 49.

[2] Admittedly, these cases are rare in the current metaethical literature. See, however, Shafer-Landau's examples of amoralism which often just work by way of the claim that in the given case, there is no relevant motivation present, and the additional claim that such cases are at least conceivable (Shafer-Landau 2003, 148–51).

[3] See, for instance, Smith 1994: 61. That there is no agreed understanding of the phenomenon of weakness of will can be seen from the fact that, unlike Smith, many metaethicists take the dissociation characteristic of weakness of will not to be one between judgement and motivation, but one between motivation and action (see Garrard and McNaughton 1998, 49; Lenman 1999, 453; and Shafer-Landau 2003, 143).

Most critics of UMI, therefore, make use of more detailed case descriptions which are meant to provide *independent* evidence for the fact that a person making a moral judgement completely lacks moral motivation. What is usually provided as evidence, however, is the failure of individuals in actual or hypothetical real life scenarios to act according to their own moral judgements (see, for example, Stocker 1979, 741–42; Dreier 1990, 10–11; Svavarsdóttir 1999, 176–77; Zangwill 2008, 102–03; and Sneddon 2009, 44–45), and evidence of this kind is necessarily inconclusive. If we take the limitation built into UMI seriously, then actual behaviour of human beings in real life scenarios cannot provide decisive evidence against UMI because it is not, and cannot be, an indicator of the absence of overridden motivation on behalf of the agent. All it can possibly indicate is the overriding motivation[4] of the agent – because it is this motivation that produces and explains the agent's action or omission. Therefore, as long as the situation in question generally allows for the agent to have more than one kind of motivation – as is the case with virtually all real life scenarios –, the agent's behaviour will tell us nothing about what is crucial for assessing UMI: whether any moral motivation was present that was too weak to produce action in accordance with the moral judgement.

Perhaps in response to this problem, some critics of UMI add a further piece of evidence to their examples: the fact that the agent himself reports of having no relevant motivation (see Svavarsdóttir 1999: 176–77; Zangwill 2008, 102–03; and Sneddon 2009, 44f.; see also Dreier 1990, 10–11; and Shafer-Landau 2003, 152). However, it is extremely doubtful whether the agent himself is a reliable judge of the matter in question.[5] Admittedly, in everyday life we often rely on the self-description of agents if we want to know what

[4] Here and in what follows, I will use 'overriding motivation' as a short hand for 'motivation sufficient to produce action in an actual situation'. In other words, the term is meant to apply not only to motivations which prove stronger than the motivations opposed to it, but also to motivations which issue in action simply because they are not opposed by any other motivations at all.

[5] For a similar worry, see Garrard and McNaughton 1998, 50.

moved them to behave as they did. Yet, we only do so because we have no choice, having no alternative access to other people's motivations. Our willingness to sometimes rely on self-reports, therefore, does not imply that we generally expect introspection to provide perfectly reliable evidence for the motivational forces within an agent. Moreover, it seems that we should expect agents to be much better able to judge their own motivations when the motivation asked for is the *strongest* motivation, that is, the one that actually led the agent to behave as he did. Yet, what we need to grasp in order to assess UMI is motivation that was too weak to issue in action, and it is especially doubtful whether self-perception is a reliable source here, given that we have no reason to assume that all cases of opposing motivations must be cases of conscious motivational conflict.

That a person fails to act in accordance with her moral judgement and additionally reports of having no relevant moral motivation, then, may well seem intuitively conceivable to some people. This does not show, however, that those people consider the complete absence of moral motivation itself conceivable, and it is only this latter question that is relevant for assessing UMI. One further reason for being sceptical is that many people's understanding of motivation seems to differ from the understanding which is at work in philosophical debates. Thus I take it that many people think of motivation as something with a certain phenomenal quality to it, as a kind of inner urge or strongly felt tendency that drags us towards or away from a certain object. However, the conception of motivation figuring in the internalism/externalism debate surely allows for motivational states that do not have this quality, and this is especially true for motivation that is too weak to issue in action. People's willingness to accept self-reports stating the absence of moral motivation, then, may also be due to the fact that 'motivation' is taken to refer to a kind of phenomenal experience that goes way beyond what the advocate of UMI is committed to.

It seems right, then, that one of the problems of the internalism/externalism debate is the notorious difficulty of proving the *absence* of (relevant) motivation (see Finlay 2004, 209). However, the conclusion to be drawn is not to simply concede this point to the externalist and to try to look for ammunition against his position elsewhere.[6] Rather, we should try to come up with a different kind of example that allows us to better distinguish between absent motivation on the one hand and overridden motivation on the other. Hume's thought experiment can serve as a blueprint for an example of this kind. The crucial lesson is that we need to take leave from real life scenarios and confront ourselves with examples which explicitly abstract from possible opposite motivations in describing the situation in which an agent acts. The behaviour of agents can give us an idea of whether or not they were motivated in the relevant sense only if the situation is one in which any relevant motivation of the agents must necessarily reveal itself in action. However, we will not encounter such situations as long as we stick to real life scenarios.

There are two further considerations to be gathered from Hume's example that also seem important for designing a reliable test case for UMI. Both considerations are meant to foreclose the possibility that, despite the explicit exclusion of any opposite motivation, the test case may still be biased towards the externalist. The first consideration is that it should be possible for the agent to perform the morally required action almost effortlessly – which makes it less plausible that the motivation to spare oneself efforts might override a present, but weak motivation to act in accordance with the moral judgement. Thus, the situation should be one in which all that is asked of the agent is, literally, to only make one move with his finger or to only speak one word, so that the motivation to act coincides almost necessarily with the action itself, and no dissociation of motivation and action is to be expected.

The second consideration concerns the relation between the content of the moral judgement and the action that is to be performed. Alleged counterexamples to UMI usually focus on cases in which the moral judgement in question prescribes a type of action as morally obligatory and in which the agent himself is to perform an act of this type. This focus

[6] This strategy is suggested by Finlay.

is understandable. Of course, internalists are committed to the claim that a person judging it morally wrong to torture other human beings is motivated to refrain from torturing human beings himself – and not only motivated to, for example, criticize those who do so. The problem, however, is that where the person making the moral judgement is himself the direct addressee of the judgement, we usually have reason to assume that there are opposite motivations that may override this person's moral motivation. What our moral obligations ask us to do is frequently, or even paradigmatically, opposed to our short-term self-interest, and this means that there will almost always be a reason to suspect that when a person does not act in direct accordance with her moral judgement, there may have been self-interested motivations at work. Yet, if this is correct, we may tend to unconsciously attribute opposite motivations to the agent even where the case description explicitly denies such motivations.

In order to avoid a bias of this kind, we should choose an example in which the action asked of the agent is more indirectly related to the moral prescription. Only if the agent fails to act morally even in those situations, it will be implausible for the internalist to claim that the agent chose the non-moral course of action only because some non-moral motivation proved stronger than his moral motivation. Of course, in doing this, we will be moving away somewhat from our own initial definition of UMI. However, it seems uncontentious that the motivations attached to moral judgements are manifold and may include the motivation to influence some other person to refrain from performing a morally wrong action. It seems legitimate, and faithful to the spirit of UMI, therefore, to enhance our initial definition and subsequently focus on the new aspect:

> *Unconditional motivation internalism** (UMI*): necessarily, if a person judges that it is morally wrong to ϕ, then she is, at least to some extent, motivated to refrain from ϕ-ing and/or to keep others from ϕ-ing.[7]

Consider, then, the following variant of Hume's thought experiment as one possible way of illustrating the true commitments of UMI: Imagine a person M who thinks that lying is morally wrong and is asked to choose between two possible states of the world (A and B) that are distinguished by one feature: while in world A, people are honest with each other, in world B, they frequently lie to each other. Imagine, further, that apart from his moral conviction, M has no incentives to choose either of the two worlds: M is not going to live in any of the two worlds himself; no person will witness and sanction his decision in any way; M does not feel compassion or other benevolent emotions towards other people; and M has no prior desire to always do what is morally right (or to always refrain from doing what is morally wrong). Now like the opponent of altruism in Hume's example, the opponent of UMI is committed to answering that M will not be able to make a decision at all because he lacks any motivation to choose one world over the other – or committed to make plausible why, despite the explicit description to the contrary, there is still some other motivation at work that allows us to expect him to choose one of the two options. In contrast, if we accept UMI, we can give what seems to be a more plausible answer: that M will choose world A because world A seems to him better than world B in virtue of being *morally* better, and because there is no motivation that could outweigh or override this consideration.

In the remainder of this paper, I will try to examine whether UMI, if explicated in this sense, is, in fact, superior to externalism or whether even for the above example an externalist

[7] Definitions of UMI that are likewise enhanced are employed by Richard Hare (see Hare 1999, 96 and 98), but also by critics of UMI, such as Jamie Dreier (see Dreier 1990, 10). Other critics of UMI provide definitions that at least suggest an enhancement of this kind (see, for instance, Svavarsdóttir 1999, 163).

indifference claim is plausible. In doing so, I will try to strictly distinguish between two meta-theoretical understandings of internalism and externalism which are both influential in the internalism/externalism debate. On the one hand, I will examine UMI for what it is usually taken to be: a conceptual claim, a claim concerning the concept of moral judgment and its implications. However, since one of the most important trends in the internalism/externalism debate is the attempt to bring empirical psychological evidence – such as evidence gathered from studies with psychopaths or so-called ‘acquired sociopaths’ – to bear on the issue in question, I will first examine UMI as an empirical psychological thesis: as the thesis that, as a matter of empirical psychological fact, a person making a moral judgement is always to some extent motivated to perform the action she holds morally required and/or to get others to perform it.

4. Unconditional motivational internalism: Psychological

It has recently been claimed that evidence from psychopaths or so-called ‘acquired sociopaths’, that is, patients with ventromedial cortex damage (hereafter: sociopaths), is pertinent to the refutation of UMI because psychopaths or sociopaths present real life examples of the key character in most externalist arguments: the amoralist, i.e. the person who is completely unmoved by her moral judgements. The strategy of internalists that try to deal with such empirical psychological findings has been to stick to the argument traditionally used against the challenge of amoralism: to deny that psychopaths or sociopaths make proper moral judgements. The idea behind this response is not that the judgements of psychopaths or sociopaths are not proper moral judgements *just because* they do not exert motivational force – which would, in principle, be a legitimate internalist reply but simply take the truth of internalism for granted. The idea is that even if, for the sake of the argument, we accept a limited or reduced conception of moral judgement that excludes the aspect of motivation, the judgements of psychopaths or sociopaths still fall short of being proper moral judgements because they lack certain formal or material features characteristic of such judgements (see, for example, Prinz 1999; and Kennett and Fine, 2008).

Now, while I generally think that there is much to be said for this view, I also think that defenders of internalism have far too generously conceded the second half of the amoralist challenge: that psychopaths or sociopaths are not motivated to act in accordance with their alleged moral judgements. If the above explication of UMI is correct, then it is far from obvious that psychopaths, sociopaths or other alleged real life examples of the amoralist lack the motivation required to vindicate UMI. What seems to make psychopaths and sociopaths an attractive piece of evidence for externalists is that they do not act in accordance with certain widely accepted moral principles. However, it should be obvious that, by itself, this well-known fact does not establish the falsity of UMI as an empirical claim. Since UMI only claims that individuals are to some extent motivated to act in accordance *with their own moral judgements*, the least the refutation of UMI requires is the demonstration that psychopaths and sociopaths fail to act in accordance with those moral principles they *themselves* endorse.

However, it is also not sufficient to show that psychopaths or sociopaths do, at some point, act contrary to moral judgements they, at some point, endorsed. What needs to be shown is rather that psychopaths or sociopaths act contrary to what has been referred to as "in situ" (Kennett and Fine, 181) judgements, that is, moral judgements that are endorsed when action according to the judgement is actually due. While it is true that the internalist claim should generally extend to other forms of moral judgement as well,[8] only a mismatch between *in situ* judgements and *in situ* action can provide the kind of empirical refutation that the opponents of UMI need. The reason is that in all other cases at least one alternative

[8] For a criticism of Kennett and Fine's restriction along these lines, see Roskies 2008, 193–95.

explanation can be given for a possible lack of motivation: that the agent altered or disowned the moral judgement he had hitherto endorsed.

Most importantly, however, even a mismatch between the *in situ* judgements and the *in situ* actions of psychopaths or sociopaths would not conclusively refute UMI because the actual behaviour of individuals will hardly ever tell us anything about the presence of overridden moral motivation. In order to empirically falsify UMI, the opponent of UMI needs to come up with an experimental setting which ensures that all kinds of motivation that could possibly override the motivation resulting from a person's *in situ* judgement are, in fact, absent. Only if, even in such a setting, a person fails to act according to a relevant *in situ* judgement, we would have reason to assume that moral judgements may entirely lack motivational force. However, it is hard to see how such an experimental setting should ever be realised.

To begin with, the setting would need to exclude the presence of any form of egoistic motivation whatsoever. The enormity of this task can be assessed by again turning to the egoism/altruism debate and, in particular, to Daniel Batson's attempts to empirically refute psychological egoism (see Batson 1991). The key problem of Batson's studies, and one which critics quickly pointed out (see, for instance, Cialdini, 1991), is that Batson relied on experiments which tested his altruistic explanation of helping behaviour, the 'empathy-altruism hypothesis', against possible egoistic explanations on a one-to-one basis. The best the experiments could show, therefore, was that test persons did, at one point, not act upon *one particular* type of egoistic motivation – which is clearly insufficient for a refutation of psychological egoism. In order to provide such a refutation, one would have to come up with an experimental setting which simultaneously tests the empathy-altruism hypothesis against all plausible egoistic explanations one can think of, and there is good reason to doubt that such a setting can be realised. However, since the motivation resulting from moral judgements can also be overridden by motivations we would not ordinarily classify as egoistic, such as motivations resulting from aesthetic feelings or even from benevolent passions, the opponent of UMI faces a challenge that is even more difficult than the one faced by the defender of altruism.

It seems, therefore, that not even psychopaths' or sociopaths' failure to act in accordance with their own *in situ* judgements will really help us to adjudicate between UMI and externalism as empirical psychological claims. The only strategy for refuting UMI as an empirical psychological claim, it seems, is to try to get a more direct access to the motivations of psychopaths or sociopaths, one that is independent of their outward behaviour. That we can directly measure the motivations of individuals is one of the crucial claims of Adina Roskies who argues that skin-conductance responses (SCRs) are a reliable indicator of motivation for action and points out that, unlike normal adults, patients with ventromedial cortex damage do not generally produce SCRs when presented with emotionally-charged or value-laden stimuli.[9] However, for a variety of reasons, Roskies' argument fails to provide a solid refutation of UMI.[10] The most serious problem, and the one most relevant for our present concern, is that Roskies does not provide good enough reasons for believing in the kind of intimate connection between SCRs and motivation that her argument relies upon. The reason Roskies views SCRs as a reliable indicator of motivation is that the presence of SCRs is reliably correlated with cases in which action is consistent with judgement, while its absence is reliably correlated with cases in which agents fail to act in accordance with their judgements (see Roskies 2003, 57). This means, however, that the reasons for taking SCRs to be a reliable indicator of motivation derive, after all, from evidence relating to the outward behaviour of individuals and, therefore, from evidence relating to their *overriding* motivations. Roskies' decision to use SCRs as an indicator of motivation is not, as it should be, based on evidence that SCRs correlate with any motivation whatsoever, but on evidence

[9] For the following, see Roskies 2003, 57.

[10] For a detailed discussion of many of these reasons, see Kennett and Fine 2008.

that SCRs correlate with motivations strong enough to produce actions. However, this means that Roskies' strategy for directly measuring motivation depends on the already refuted possibility of inferring overridden motivation from actual behaviour.

The above considerations not only question the success of Roskies' specific attempt to refute UMI, but seem to have implications for any attempt of this kind. It seems that any experimental strategy for directly measuring motivation must run into the same kind of problem. Thus it is hard to see how we could ever produce evidence that certain physiological phenomena – be it SCRs, brain region activities, or something else – are strictly correlated with motivation other than inferring this from the way these phenomena correlate with human behaviour. The reason is that human behaviour is the only observable phenomenon that the concept of motivation is undoubtedly associated with. The concept of motivation is as complex and disputed as the concept of moral judgement, and the only thing that seems uncontentious is that the concept of motivation is related to the concept of action in a way that excludes the possibility of unmotivated action. However, if, in searching for motivational evidence, all we can go by is the fact that every human action coincides with motivation, it seems that all we will ever be able to safely identify is the presence or absence of the kind of motivation that issues in action, and this is overriding motivation. To be able to empirically identify and measure overriding motivation, however, is not going to allow us to refute UMI as an empirical claim.

The first conclusion to be drawn, then, is quite strong. It is not only that the strategies hitherto used to refute UMI as an empirical psychological claim provide us with inconclusive evidence. It is also that any strategy to refute UMI as an empirical psychological claim will necessarily face similar problems. As an empirical psychological claim, then, UMI might generally withstand refutation. However, the reason for this is not, as one might have thought, that the defender of UMI can always deny that an alleged real life example of the amoralist makes genuine moral judgements. Even if one is prepared to make this concession, UMI seems to be well guarded against purported empirical evidence in favour of externalism. What remains to be seen, then, is whether there is a way to adjudicate between UMI and externalism as conceptual claims, and whether empirical evidence can play any significant role in this.

5. Unconditional motivational internalism: Conceptual

Our conclusions regarding UMI as an empirical psychological claim do not, in any relevant sense, establish a prejudice with regard to the assessment of UMI as a conceptual claim. Obviously, both the truth and the falsity of the conceptual variant of UMI would be compatible with the conclusion that UMI may not be refuted as an empirical psychological claim. However, not only the fundamental question of whether UMI as a conceptual claim is true or false remains an open question. Just as open are the questions of whether we can *prove* the truth or falsity of the conceptual variant and whether we can do so by means of empirical evidence. That the latter possibility still exists is due to the fact that the evidence needed in order to refute UMI as an empirical psychological claim concerns people's actual moral judgements and their actual motivations. Yet, it seems quite clear that this kind of evidence will not tell us anything about the characteristic features of our concept of moral judgement – even though it might give us reasons to rethink and to revise our concept. If there is an empirical way of finding out something about these features at all, it seems that the evidence we should go for is linguistic or anthropological rather than psychological.

One key question in trying to determine the truth or falsity of UMI as a conceptual claim, therefore, is whether we generally think that non-psychological empirical evidence, such as linguistic evidence, can tell us anything about our concepts, and we seem to have good reasons for thinking that it does. The least we should assume is that the ways in which competent speakers use terms that are intimately connected with a certain concept are *indicative* of that concept. The same applies to the views ordinary speakers hold with regard to those terms. Following an influential approach in meatethics, we may, for instance, think that folk platitudes surrounding a specific concept may help us to describe the concept's characteristic features (see, for example, Smith 1994; Jackson/Pettit 1995; and Jackson 1998), and we may assume that these folk platitudes, in turn, can best be identified by way of empirical investigation. *Prima facie*, therefore, it seems that one thing to be done in order to overcome the impasse of philosopher's intuitions regarding the concept of moral judgement is to collect data concerning the actual usage of relevantly related terms or relevantly related folk platitudes.

In order to consider this a viable and fruitful approach, we need not assume that we can simply read off the concept of moral judgement from actual ethical discourse, or that actual usage is all there is to the concept of moral judgement. For instance, we can expect ordinary speakers' assumptions and linguistic habits to be somewhat incoherent. Determining the features of moral concepts, then, is not just a matter of collecting ordinary speakers' opinions and linguistic habits, but a matter of critically reconstructing them. Still, it seems that if we entirely lack such an empirical basis, we will not be able to make any headway concerning the conceptual question at all but will be stuck with philosophers' intuitions about the concepts in question, such as, for instance, their diverging intuitions about whether or not counterexamples to UMI are possible or conceivable.

There have hitherto only been few attempts to collect ordinary speakers' relevant moral platitudes or to describe their use of relevant moral terms (see, however, Nichols 2002 and, more recently, Strandberg/Björklund, forthcoming, and Björnsson et al., forthcoming). One reason why statistical approaches have hardly been pursued may be that UMI seems too sophisticated a claim as to be settled by collecting people's opinions about it. It is true that most metaethical questions can hardly be resolved by simply posing them to ordinary speakers – primarily because much of the knowledge relevant for answering the questions is implicit rather than explicit knowledge. However, while it may follow from this that statistical approaches to metaethical questions face certain problems, a general rejection of such approaches is uncalled for. The conclusion to be drawn is rather that, in order for surveys and other empirical methods to yield reliable results, special care is needed. Trivially, of course, one should generally strive to avoid terminology that is too technical or that is ambiguous and hence likely to give rise to confusion. In addition, one should strive to come up with test questions that are capable of revealing relevant implicit knowledge without directly asking for it.

Now, as a matter of fact, Hume's thought experiment gives us an example of what such an indirect strategy might look like. It does not directly ask us for our views about motivation and thus avoids problems that might result from the fact that ordinary speakers may be unsure about what exactly is meant by 'motivation' or may not be conscious of all the concept implies. Still, it is capable of revealing ordinary speakers' views about what forms of motivation we would expect to find in ordinary human beings. The same seems to apply to our variant of Hume's thought experiment. By focusing on the aspects relevant for assessing UMI's motivational claim and, at the same time, abstracting from irrelevant aspects that may introduce biases, the example provides us with the opportunity of collecting ordinary speakers' views about the connection between moral judgement and motivation without directly asking for those views. Therefore, our variant of Hume's example suggests that the debate over UMI as a conceptual claim may be resolved by way of empirical methods because the example itself suggests an empirical approach for identifying the relevant aspects of our concept of moral judgement.

One may object here that, given the pitfalls of linguistic surveys, even an indirect strategy will only provide us with evidence that is all as unreliable as the physiological and psychological evidence we have criticized above. Now, it is true that results from linguistic surveys remain challengeable and that we must expect at least one party to the internalism/externalism debate to maintain a general scepticism with regard to our results. However, the psychological and the linguistic cases are still different in an important sense. In the linguistic case, it seems uncontentious that the data we may obtain with the help of surveys represents a relevant kind of evidence because it seems undeniable that the answers given are in some way controlled by the test person's concepts and hence indicative of those concepts. However, the psychological case is different. As has been argued above, actions in real life scenarios are not indicative of overridden motivation, and the question of whether certain physiological reactions are indicative of overridden motivation inevitably seems to remain open. Though both concerned with empirical evidence, the two cases are therefore characterised by an important asymmetry. The problem in the linguistic case is that there may be biases or side effects that affect our results. Yet, unless there are good reasons to assume that there were, in fact, such biases or side effects in a given case, any speculation that the evidence is unreliable appears *ad hoc*. The problem in the psychological case, in contrast, is that we cannot be sure that the obtained evidence has anything to do with what we want to examine at all. The refusal of purported evidence against UMI as irrelevant or inconclusive, therefore, is not *ad hoc* in the same sense. We may also put this point in terms of the burden of proof: in the linguistic case, the burden of proof lies with the one who questions the reliability of the results since he claims that what is generally an indicator of concepts does not provide reliable evidence in the given case. In the psychological case, the burden of proof lies with the one who presents his empirical results as reliable because he must provide reasons for thinking that he has identified an indicator of overridden motivation in the first place. Although both approaches may be open to objections, then, the linguistic approach seems in a stronger position because it starts from the shared assumption that what people say, and what they take other people to say, has to do with the concepts they employ.

Furthermore, it is important to emphasize that a survey based on our variant of Hume's thought experiment seems much better suited to reveal the truth about UMI's motivational claim than the few other empirical studies that have hitherto been conducted, such as the recent study of Shaun Nichols.[11] In Nichol's study, philosophically unsophisticated undergraduates were presented with the example of John, a psychopathic criminal, who agreed that hurting others was morally wrong, but confessed that he did not care whether he did things that were wrong, and who had actually killed other people. Students were then asked whether they thought that John really understood that hurting others is morally wrong. According to Nichols, nearly 85% of the students maintained that John did really understand this. In his final summary, Nichols concludes from this that "psychopaths are commonly regarded as rational individuals who really make moral judgments but are not motivated by them" (Nichols 2002, 301).

It is doubtful, however, whether the design of Nichols' survey really supports this conclusion. As we have shown, in order to assess UMI we need to come up with examples or thought experiments that allow us to discriminate between absent motivation on the one hand and outweighed or overridden motivation on the other. Yet, Nichols' example, which focuses strongly on John's moral judgement and treats the complex aspect of motivation only as some sort of appendix, clearly fails in this respect. To begin with, Nichols' study did not test students' intuitions about whether John lacks moral motivation at all: all the students were asked was whether they thought that John really understood that hurting others is morally wrong. Nichols simply takes it for granted that, besides thinking that John really understood that hurting others is wrong, students also thought that John completely lacked moral motivation. However, Nichols does not seem entitled to conclude that, and not only because the students were not given any formal opportunity to contradict this aspect of the example. Nichols' description just does not make it sufficiently clear to the reader that moral motivation is completely absent from John. First, it only cites John's self-report as evidence –

[11] For a detailed critical discussion of Nichol's study, see also Björnsson, forthcoming.

"He *says* that he knows that hurting others is wrong, but that he just doesn't care if he does things that are wrong." (Nichols 2002, 289, my emphasis) – which report a reader may or may not find reliable. Secondly, it is not obvious that John thereby means to report a complete absence of motivation: neither can we just presuppose that the concepts of caring for something and of being motivated to do it are identical, nor can we assume that the difference between absent and overridden motivation is clear enough to philosophically unsophisticated undergraduates as to make them understand John's remark as referring to the former, but not to the latter.

It seems, therefore, that our variant of Hume's thought experiment should be a better guide to ordinary speakers' views about the possibility of amoralism than the example used by Nichols. In light of these considerations, I conducted surveys at the universities of Cologne, Aachen and Münster in which I used the thought experiment as a means of collecting ordinary German speakers' views about the relation between moral judgement and motivation. About 120 undergraduate and graduate students were each presented with one of four different versions of the thought experiment, the first and most straightforward version being the following:

> Imagine the following situation: A person, let's call him M, sincerely thinks that it is morally wrong to lie to other people. By chance, M is able to determine, just by pressing a button, how individuals on a distant, recently populated planet are going to behave in the future. If M presses button A, the individuals will be honest with each other. If M presses button B, many of them will repeatedly lie to the others. What will M do? (version 1)

The example was supplemented by one of two following sets of *Further assumptions* about M:

> M himself is never going to live on planet P; nobody will witness his decision or find out about it afterwards; M generally does not feel compassion towards other people (set 1)
>
> M himself is never going to live on planet P; nobody will witness his decision or find out about it afterwards; due to certain psychopathological disorders, M is unable to feel compassion towards other people (set 2).

The students were then asked to choose one out of four possible answers:

> … M will press button A.
>
> … M will press button B.
>
> … M is unable to make a decision at all.
>
> … What M does depends on additional factors, such as………..

The answers given by the students provide important support for UMI as a conceptual claim. Of the test persons presented with version 1, over 80% predicted that M would press button A and thereby ensure that people on planet P would be honest with each other. Less than 13% claimed that M would be unable to make a decision, while the remaining 7% thought that M's decision was dependent upon further factors. These results, which were not significantly affected by whether test persons were given set 1 or 2 of the *Further assumptions*, suggest that a large majority of ordinary speakers may hold an internalist view of moral judgement, and, in fact, an unconditional internalist view. The least the results indicate is that the majority of our test persons takes moral judgements to be potentially motivational, or sufficient to give rise to motivation by themselves. However, since it has recently been argued that mental states may be potentially motivational without being necessarily motivating (see Shafer-Landau 2003, 147–48), one may still think that the results fall short of providing evidence for UMI. Yet, it seems that the distinct and stronger claim that moral judgements necessarily motivate (if it is one) is supported by our results as well. After all, test persons who thought that M might press button A, but would perhaps not do so on every occasion, were given the opportunity to communicate this by answering that what M does depends on further factors. There seem to be no reasons to assume that undergraduate students are incapable of understanding the difference between answers A and D or of connecting it to their predictions about what M is going to do.

Still, the defender of externalism may find reason to object to the above conclusions. He is likely to argue that test persons may have attributed to M an independent desire to do the morally right thing, and likely to support this supposition by claiming that we generally expect some such desire in other individuals, or at least in those individuals who have clearly specified moral views. This response should seem a bit *ad hoc*, though. It needs to be emphasized that the case description does not positively attribute to M any prior and independent desire to do the morally right thing and that test persons were explicitly asked to stick to the given description and not add any further assumptions of their own. Moreover, we may well want to challenge the underlying claim that we generally expect other individuals to have an independent desire to do the right thing. Yet, given that the idea behind our survey was to use case descriptions that explicitly abstract from all relevant motivations which do not directly result from the moral judgement itself, it seemed that we should try to strengthen the example by giving test persons even less reason to assume that M could be acting out of a prior and independent desire to do what is morally right.

The following alternative versions were shaped by this consideration:

> Imagine the following situation: By chance, a person, let's call him M, finds himself able to determine, just by pressing a button, how individuals on a distant, recently populated planet are going to behave in the future. If M presses button A, the individuals will be honest with each other. If M presses button B, many of them will repeatedly lie to the others.

> (version 2a) M compares the two scenarios and, in doing so, he contemplates, for the first time in his life, the moral qualities of actions. In the course of his reflections, he comes to the conclusion that, as a matter of principle, it is morally wrong to lie to other people. What will M do?
>
> (version 2b) M considers what reasons might speak in favour of one of the two scenarios, but cannot come up with any such reasons at first. Moral matters have never interested M so far. However, he has casually noticed that some people consider lying a morally relevant issue, and so he finally asks himself whether there might be moral reasons that speak in favour of either of the two scenarios. M comes to the conclusion that, as a matter of principle, it is morally wrong to lie to other people. What will M do?
>
> (version 2c) M considers what reasons might speak in favour of one of the two scenarios, but cannot come up with any such reasons at first. Moral matters have never interested M so far. However, he has casually noticed that some people consider lying a morally relevant issue, and so he finally asks himself whether there might be moral reasons that speak in favour of either of the two scenarios. M comes to the conclusion that, as a matter of principle, it is morally wrong to lie to other people and that by pressing button B, he would bring about a morally worse state of affairs. What will M do?

In all three cases, M is presented as an individual that has not thought much about morality so far and has not himself striven to evaluate human behaviour in terms of right and wrong at all. This seems to deprive the assumption that M has a standing desire to the morally right thing of its basis: if a person has a standing desire to do what is morally right, we would expect her to also have an instrumental desire to find out what is morally right. Moreover, we might expect the standing desire to do the right thing to be itself a result of (primitive) moral reflection, which gives us even more reason not expect such a desire in a person that has never contemplated moral issues at all.

The answers given by test persons confronted with these examples indicate that the modifications have some impact on peoples' considerations. However, they did not significantly affect the general internalist tendency of the predictions. In the case of version 2c, this tendency was even stronger than in the first example. Thus more than 83% of the test persons thought that M is going to press button A, while the other three answers were each chosen by less than 6% of the test persons. In contrast, test persons confronted with version 2a and version 2b still predominantly chose answer A, but they did so less often than those confronted with version 1. In the case of version 2a, predictions that M will press button A went down to 59%; in the case of version 2b, they went down to 69%. In both cases, the number of test persons that opted for the indifference claim (answer C) remained relatively constant (15% in the case of version 2a, 14% in the case of version 2c). The number of participants that claimed that M's decision depends on further factors went up slightly, to 10% in both cases. We may note, then, that the prediction that M will press button A is, in two cases, negatively affected by our modification of the original example. In both cases, however, it remains by far the most popular answer, the percentage of predictions that M will press button A remaining as high as 71% overall.

Notwithstanding the variations in the responses to the different versions of the questionnaire, therefore, the results of our surveys speak in favour of UMI as a conceptual claim. Obviously, however, they do not allow us to straightforwardly claim its truth or to conclude that any critical reconstruction of the concept of moral judgement based on linguistic usage and folk platitudes needs to include UMI. There are two questions that remain to be addressed. The first is how the evidence for UMI as a conceptual claim is to be weighed against the counterexamples discussed in the literature which are at least considered intuitively compelling by their externalist creators. The second question is what we should make of the fact that even in our own survey, some test persons did not display an internalist view of moral judgement, but opted for an externalist answer. The general underlying question is how to deal with conflicting evidence concerning our moral concepts: is there still a way to claim that one of the two conceptual claims discussed in the internalism/externalism debate is the correct one, or are we left with the conclusion that either both claims are true or neither of them is?

The first question does not seem to pose any serious problems. In criticising the previous attempts to refute UMI as a conceptual claim, I raised general doubts as to whether the supposed counterexamples do justice to UMI's motivational claim. Now if one follows this line of thought, one seems to have good reasons for thinking that the evidence from our surveys simply trumps the available counterexamples to UMI. This claim can be backed up by two further considerations. The first is that the counterexamples launched against UMI can be given, and have been given, other plausible internalist interpretations. The internalist need not claim that the counterexamples previously discussed in the literature fail to ensure the complete absence of moral motivation. He can also claim that the alleged amoralists do not make any genuine moral judgements. One may find this interpretation *ad hoc*. However, it seems that the externalist will have a much harder time explaining away the fact that such a large majority of test persons expects the person in our thought experiment to ensure honesty on planet P by pressing button A. It seems far easier, then, to reinterpret the available evidence for externalism in an internalist fashion than to reinterpret our evidence for internalism in an externalist fashion.

Secondly, it is not at all clear whether the alleged conceivability of the amoralist would really put UMI in jeopardy. In order to show that no necessary link between judgement and motivation is comprehended in our concept of moral judgement, the bare fact that we can *imagine* amoralists does not seem sufficient. It would be sufficient if our concept of moral judgement were as clear-cut and dry as the notorious ‘bachelor’-example. In the case of ‘bachelor’, we may indeed argue that a ‘married bachelor’ is not even as much as imaginable or conceivable because any competent speaker would instantly perceive the conjunction of subject and predicate to be inconsistent. However, it seems indisputable that most of our moral concepts, including the concept ‘moral judgement’, are much more complex than ‘bachelor’ and have many implications that are far from being obvious. That conceptual questions have given rise to so many long-standing debates in metaethics – debates that are unheard of with regard to ‘bachelor’ – seems to provide ample evidence for this. However, if the manifold implications of our moral concepts are not obvious to ordinary speakers all the time, we may use this fact in order to question the force of the alleged conceivability of the amoralist.

In light of these considerations, it can be argued that we have two plausible strategies for dealing with conflicting linguistic evidence concerning moral concepts. The first is to point out that ordinary speakers may sometimes be unfaithful to their concepts, especially with regard to somewhat hidden implications of these concepts. Although a certain implication may generally be part of a concept, a speaker possessing this concept may at times accept propositions which are incompatible with this implication because he is not fully aware of the incompatibility of the two. A second, related strategy is to claim that the conceptual views or linguistic habits of a speaker may generally be shaped by assumptions that are incompatible. Applied to the case in question, we may assume that people who accept the externalist counterexamples as plausible may only do so because they fail to perceive that the existence of the amoralist is incompatible with assumptions about the connection between moral judgement and motivation they themselves hold. However, given the problems of the alleged externalist counterexamples we have already described at length, we seem to have good reason to argue that a critical reconstruction of the concept of moral judgement based on linguistic evidence should include the assumption that moral judgements necessarily motivate rather than the rival assumption that amoralists are possible.

The two strategies sketched above provide important alternatives to a third one: that of conceding that there is no such thing as *the* concept of moral judgement because different speakers simply employ different concepts. It seems, however, that when it comes to dealing with the second of the two questions mentioned above, this third option may be the only one we are left with. Thus it seems that the first two strategies cannot explain why in our own survey, some test persons gave an externalist rather than an internalist answer – or only explain it in a way that would undermine our other conclusions. It seems, then, that one of the results of our inquiry is that ordinary speakers' concepts of moral judgement differ – at least with regard to one aspect, namely the connection between moral judgement and motivation. While most people seem to hold the unconditional internalist view that a person making a moral judgement is necessarily motivated to act in accordance with this judgement, others apparently think of the link between judgement and motivation as contingent.

Our conclusion regarding UMI as a conceptual claim, then, needs to be a qualified conclusion. The results of our survey do not suggest that UMI is generally true. They only suggest that UMI does capture a feature of most ordinary speakers' concept of moral judgement, capture it, that is, better than externalism and weaker versions of internalism. If nothing else, this conclusion is at least sufficient to prove the overall point with which we started: that the widespread rejection of UMI is premature and that UMI deserves serious consideration. However, a further tentative conclusion to be drawn from our survey is that, even if we strive harder to bring out the true commitments of UMI and to test it for what it really is we will still not be able to fully overcome the impasse of intuitions that has long characterized the internalism/externalism debate. This impasse seems to have at least some foundation in the fact that ordinary speakers' concepts of moral judgement motivation differ.

If this is correct, however, then the truth of internalism may be that, depending on how we look at it, neither internalism nor externalism is true or both are.

For valuable comments and suggestions I am grateful to Ragnar Francén Olinder, John Eriksson, Gunnar Björnsson, Jesse Prinz, Kate Mann, Wilfried Hinsch, Markus Stepanians and Anna Maj Blundell. I would also like to thank the students who participated in my surveys and the colleagues who allowed me to hand out my questionnaires in their lectures: Kurt Bayertz, Oliver Scholz, and Thomas Gutmann. Finally, I would like to express my gratitude to the *Deutsche Forschungsgemeinschaft* (German Research Council) for funding the project on expressivism and moral sentimentalism (HI 376/6-1) from which this paper has emerged.

References:

Batson, C. D. 1991. *The altruism question: Toward a social psychological answer.* Hillsdale, NJ: Erlbaum.

Björnsson, G. et al. “Motivational internalism and folk intuitions.” *Philosophical Psychology*, forthcoming.

Cialdini, R. B. 1991. “Altruism or egoism? That is (still) the question.” *Psychological Inquiry* 2: 124–126.

Dreier, J. 1990. “Internalism and speaker relativism.” *Ethics*. 101: 6–26.

Finlay, S. 2004. “The conversational practicality of value judgement.” *Journal of Ethics* 8: 205–223.

Francén, R. 2010. “Moral motivation pluralism.” *Journal of Ethics* 14: 117–148.

Garrard, E., D. McNaughton. 1998. “Mapping moral motivation.” *Ethical Theory and Moral Practice* 1: 45–89.

Hare, R. M. 1999. “Internalism and externalism in ethics.” In R. M. Hare *Objective prescriptions and other essays*, Oxford: Clarendon Press, 96–108.

Hume, D. 1777. “An enquiry concerning the principles of morals.” In D. Hume *Enquiries concerning human understanding and concerning the principle of morals*, ed. by L. A. Selby-Bigge and P. H. Nidditch, Oxford: Clarendon Press, 169–323.

Jackson, F., and P. Pettit. 1995. “Moral functionalism and moral motivation.” *Philosophical Quarterly* 45: 20–40.

Jackson, F. 1998. From metaphysics to ethics: A defence of conceptual analysis. Oxford: Oxford University Press.

Kennett, J., and C. Fine. 2008. “Internalism and the evidence from psychopaths and ‘acquired sociopaths’.” In *Moral Psychology. Vol. 3*, ed. by W. Sinnott-Armstrong, Cambridge: MIT Press, 173–190.

Lenman, J. 1999. “The externalist and the amoralist.” *Philosophia* 27: 441–457.

Nichols, S. 2002. “How psychopaths threaten moral rationalism: Is it irrational to be amoral?” *The Monist* 85: 285–303.

Prinz, J. 2006. “The emotional basis of moral judgements.” *Philosophical Explorations* 9: 29–43.

Roskies, A. 2003. “Are ethical judgements intrinsically motivational? Lessons from ‘acquired sociopathy’.” *Philosophical Psychology* 16: 51–66.

Roskies, A. 2008. “Internalism and the evidence from pathology.” In *Moral Psychology. Vol. 3*, ed. by W. Sinnott-Armstrong, Cambridge: MIT Press, 191–206.

Shafer-Landau, R. 2003. *Moral realism. A defence.* Oxford: Clarendon Press.

Smith, M. 1994. *The moral problem.* Oxford: Blackwell.

Sneddon, A. 2009. “Alternative motivation: A new challenge to moral judgement internalism.” *Philosophical Explanations* 12: 41–53.

Stocker, M. 1979. “Desiring the bad: An essay in moral psychology.” *The Journal of Philosophy* 76: 738–753.

Strandberg, C. 2012. “Expressivism and dispositional desires.” *American Philosophical Quarterly* 49: 81–91.

Strandberg, C., and F. Björklund. “Is moral internalism supported by folk intuitions?” *Philosophical Psychology*, forthcoming.

Svavarsdóttir, S. 1999. “Moral cognitivism and motivation.” *The Philosophical Review* 108: 161–219.

Zangwill, N. 2008. “The indifference argument.” *Philosophical Studies* 138: 91–124.