Ratio (new series) XI 1 April 1998 0034–0006

WHAT'S WRONG WITH MORAL INTERNALISM

Robert Lockie

Abstract

Moral Internalism is the claim that it is a priori that moral beliefs are reasons for action. At least three conceptions of 'reason' may be disambiguated: psychological, epistemological, and purely ethical. The first two conceptions of Internalism are false on conceptual, and indeed empirical, grounds. On a purely ethical conception of 'reasons', the claim is true but is an *Externalist* claim. Positive arguments for Internalism – from phenomenology, connection and oddness – are found wanting. Three possible responses to the stock Externalist objections are uncovered and overturned. In so doing a close relation between Internalism and Behaviourism is revealed, and some stock anti-behaviouristic arguments are co-opted for Externalism. The likely dependence of Internalism on an Atomistic Associationism is uncovered and criticised. Internalism is seen as being ultimately a type of Ethical Determinism. Finally, a sketch of an Anti-Associative Externalism is given whereby the notion of *self determination* of action is put forward as an account of moral motivation fit to resist both the internalist and the belief-desire psychology premises of the stock non-cognitivist argument.

I

Moral internalism is the claim that it is an a priori truth that moral beliefs are reasons for action. This is a generic version of internalism, with 'belief' a place-holder for whichever term for a mental state the internalist in question prefers to characterise human moral psychology – beliefs proper, desires, 'conceptions of the circumstances', 'deliverances of the sensitivity', attitudes, sentiments, or other.

Moral internalism matters, not least because a variety of it serves as premise in all of the many versions of the following argument.

Premise One: Moral 'beliefs' alone, as a matter of conceptual necessity, provide reasons for action.
Premise Two: A reason for action may be attended by one or more

beliefs proper, which aim to represent the world and can alone be true or false; but must comprise at least one *desire*, which cannot. *So:* Moral 'beliefs' cannot be true or false.

One becomes a *cognitivist* by resisting the conclusion of this argument in some way. The *externalist* holds that Premise One is false: it is at least conceptually possible that one might have a moral belief, yet not thereby have a reason for action. The internalist cognitivist can only dispute the truth of Premise Two, and can do so on any of a number of grounds, but what is often overlooked is that the externalist may also choose to oppose this premise.

This is important because internalists often level a criticism against externalists which could only have application to those who accepted Premise Two – and a renowned externalist who has is Phillipa Foot. The criticism, in Kant's terms, is that the externalist wrongly treats the dictates of morality as if they were hypothetical not categorical imperatives: hypothetical on our possessing the further [neo-Humean, non-cognitive] *desire* to act on our moral cognitions. The externalist who objects to both premises will not see any such desire as being required for one to have a reason for action, and to labour criticisms of Premise Two against such an externalism is to push against an open door.

My own grounds for rejecting Premise Two will be given later. It suffices here to say that one need not oppose internalism under the false assumption that this is the only way to resist such arguments and remain a cognitivist. In fact, as it is commonly given, it is questionable whether the argument is even *valid*. Notably, it may be committing a fallacy of equivocation in the sense attached to 'reasons for action' across Premises One and Two. Important to what follows, 'reasons for action' can be disambiguated in at least the following three ways – and perhaps others besides – of which only the first two are commonly distinguished in the internalist literature (though less commonly kept as distinct).

(a) Psychological motivator of action: disposition to behaviour. (N.B. Henceforth we will use 'action' and 'behaviour' interchangeably, reading the internalist's qualification '. . . *or motivation* . . ., *disposition* . . ., *tendency* . . . *to*', where needed.)

(b) Normative-epistemological, that is, *rational*, grounds for action.

(c) Normative, but purely *ethical* grounds for action. *Moral* obligation to act.

Pretty clearly, Premise Two is using (a), whereas much of the time, versions of Premise One appear not to be – or not only to be. One might expect the internalist *cognitivist* to welcome this point about the ambiguity of 'reasons . . .' – for in resisting non-cognitivism, it gives a second string to her bow. I will suggest below why such a welcome has not been forthcoming. For the most part, although internalists of both a cognitive and a non-cognitive stamp are chronically vague and inexplicit between their use of (a) and (b), it is clear that (a) is at least a major aspect of internalism as such, and it is with this disambiguation of 'reasons' that we will begin considering the positive arguments for internalism, later to consider (b) and (c) in turn.

II

One may discern three stock arguments for internalism in the literature. These we can label the argument from *Oddness*, the argument from *Phenomenology*, and the argument from *Connection*.

The argument from oddness, is that in reflecting upon the externalists' alleged conceptual possibility, of someone who has a moral belief yet doesn't thereby have a reason for action, we are forced to conclude

> it would be *odd* for someone to say 'this action is wrong but I don't see that as at all relevant to my choice'[1].

The argument from phenomenology, is that by reflecting on our experience of moral belief we will be led to realise that such is essentially motivating.

> [W]hen we consider perception of *moral* value, I think it clear that, if introspective phenomenology can be our guide, moral perceptions manifest a unity of such a kind that potential motivation is indeed internal to them.[2]

The argument from connection is the positive obverse of the negative argument from oddness. We must explain the fact, so obvious that it is easily taken for granted, that moral

[1] Jonathan Dancy, *Moral Reasons* (Oxford: Blackwell, 1993), p.4. This, as with each of these arguments, has multiple locations in the literature. I should add that Dancy sees his position as ultimately only having 'an internalist flavour'.

[2] Mark Platts, 'Moral Reality and the End of Desire', in Mark Platts, ed., *Reference, Truth and Reality* (London: Routledge and Keegan Paul, 1980), p.81.

beliefs are commonly and non-coincidentally *connected* with moral actions.

> [A] *change in motivation* follows reliably in the wake of a *change in moral judgement*, at least in the good and strong-willed person. . . . As I see it . . . [internalists] can, whereas strong externalists cannot, explain this striking fact.[3]

Let us consider the *connection* argument first. Of course changes in motivation can be reliably predicted from changes in moral beliefs (judgements). Surely there is no great mystery as to why this is: changes in beliefs *cause* changes in motivation, with causation as external a connection as there is. This appears to make the connection defeasible metaphysically as well as conceptually, which most externalists take to be a virtue of their position, since it makes possible what is indeed sometimes actual: the failure to be motivated to act on what we nevertheless know to be right. (It would, moreover, anyway still be open to externalists to see the connection as necessary a posteriori – though it is hard to see why they would want to in the general case).

This response to the argument from connection carries over naturally to its obverse: the argument from oddness. Mostly, it would be *odd* to fail to be motivated to act on ones moral beliefs – but because it would be a somewhat odd (unusual) causal situation, not because it would be logically odd. Indeed, there is no such thing as *logical* 'oddness'. A married bachelor is not an odd state of affairs, it is an incoherence – not a description of a state of affairs at all. That is because, contra the moral case, 'bachelor' and 'married man' *are* internally connected.

Actually, such force as the oddness argument has, it probably has in virtue of presenting us with our first example of internalism tacitly sliding between different senses of 'reasons for action'. We listed above, three possible disambiguations of 'reasons. . .', of which only reasons (c) – *moral* reasons – makes the central claim of the argument from oddness at all plausible. It does seem fair to say that it would be odd, indeed *internally incoherent*, for someone to say:

> 'this action is [MORALLY] wrong but I don't see that as [MORALLY] relevant to my choice.'

[3] Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994), p.71.

Undoubtedly moral beliefs are *morally* relevant, and represent *moral* reasons for action – even a wholesale moral sceptic, even a knowingly wicked person, would not count against such a claim. Such figures, if they existed, would be sceptical, knowingly wicked, etc., in virtue of failing to be moved to action by such purely moral reasons, and would probably further deny they were irrational in being unmoved. It is the conceivability of being unmoved by moral reasons in these two senses – reasons (a) and possibly (b) – which internalism is committed to opposing. Yet if being thus unmoved were *inconceivable*, it would be hard to see how the notion of 'having a [moral] reason' not to be conatively unmoved would make sense. It cannot be said that one *ought* either not to 'do' the inconceivable or to do the inconceivable-not. Perhaps internalists will, when this consequence is pointed out, wish to embrace it – 'there is one thing about which it does not make sense either to say that it is a moral reason or that it is not a moral reason for action, and that is the possession of a moral belief.[4] However they wished to respond, if there were a dissatisfaction with the oddness of someone having a moral belief yet not a *moral* reason to act on it, this dissatisfaction would not be with any species of externalism here described – arguably, it would be a dissatisfaction with internalism.

Finally, we come to the argument from phenomenology. Introspecting on our own case is supposed to establish that we would find it inconceivable to possess a moral belief and not be motivated to act on it. Of course some of this phenomenological certainty may be an artefact of the same confusion as that just considered: introspecting on one's *moral* beliefs, thereby to appreciate one's *moral* reasons (reasons why one *ought to act*). The thesis advanced here must be that one can phenomenologically apprehend the certainty of one's psychological disposition to action.

To this it should be countered that introspection is an enormously implausible way to seek to establish a theory of action. A theory of action is typically about the connection between motivators in the agent and action in the world. It is highly questionable whether introspection gives reliable access to the former,

[4] Paraphrasing Wittgenstein's metre-rule example, remark 50 in the *Investigations*. For an effective criticism of such neo-operationalism in its original domain of application see Jerry Fodor and Charles Chihara: 'Operationalism and Ordinary Language', *American Philosophical Quarterly*, Vol. 2, No. 4, (1965).

and unquestionable that it fails to give reliable access to the latter, or the connection between the two. The whole of modern psychology, since its academic (experimental) and clinical (post Freud) break with the philosophical tradition of introspection, is predicated on the simple truth that most things about mind and behaviour have to be *found out* in controlled and indirect ways.

Additionally, I don't here even agree with the internalists about the phenomenology. On the few occasions when I believe I have conducted myself well, I don't recall any inner thrill of moral comprehension push-pulling me inexorably towards The Good As I Saw It. On the contrary, it seems to me from now, that it seemed often, on those occasions, that I apprehended the right way to go, and coldly, with a kind of dissociation, acted despite the felt possibility of not acting. Ignoring the Good whilst still perceiving it as the Good, was felt as very much a live option, and being *righteous* involved steeling myself against that option. Internalism doesn't seem to me, to capture the *sang froid*, deliberation, even (at times) *resentment* often involved in acting well – the taste of blood and ashes in the mouth.

Now this may just mean that there are individual differences in the phenomenology. And my phenomenological introspection may be in error even as it applies to myself (though I assure you it is ingenuously put forward). But that just points up the fallibility of phenomenological introspection even as *phenomenological* introspection. As data for a substantive claim about the springs of action, one might doubt whether it yields the truth much above levels predicted by chance.

III

The possibility that moral judgement and motivation to action could be dissociated, really just is an inchoate statement of externalism, so it is to be expected that the stock objections to internalism develop from this. The stock objections derive from considering the possibility of evil. After Dancy, we can label them the problems of *Wickedness, Amoralism*, and *Accidie*, though the divisions are somewhat arbitrary. As we shall see, it is evil tout court that is the real objection.

The wicked person is defined as a strict moral invert. Indeed, as internalists (suspiciously) gloss the wicked man, he is a confirming instance of internalism, but *inverted* internalism. He has the cognitive ability to discriminate immorality correctly

enough, yet is held to be dissociated from the normal response to that immorality by being attracted to evil in itself. The amoralist is not especially attracted to immorality for its own sake, but will act disregarding his accurate moral cognitions, or rather, indifferently to them. The sufferer from *accidie* is temporarily separated from an otherwise normal direction of moral motivation – life events, *ennui*, etc., lead her to be misanthropically dissociated (for a time) from any motivation to act morally on what are still accurate moral cognitions.

Actually, to this externalist demonology should be added an optimistic case. The dissociation could go in the other direction. One might accurately divine that X was right, and act in the direction this cognition points, yet with one's motivation split from that moral belief, deriving from, or mediated-by, another source – perhaps redoubled by that source. As when agents act via high-minded general moral principles and theory, or from the attempt to be consistent, or from a patrician sense of 'noblesse oblige', or from the same acidic *ennui* (product of a broken heart, say) that in the pessimistic case produced a misanthrope or worse; but here is a source for good – sending forth the burnt out case to toil in the leprosery. Moral monsters and moral saints are not always so far apart, one might think. The difference each has from the rest of us is a pathological source of motivational energy, the similarity, is a good-enough ability to correctly discriminate and categorise different moral cases.

In each externalist case, the belief that X is right is held not to motivate *simply*, *immediately*, *internally*, in virtue of being that belief, X. The realisation that this is a possible cognitivist account of moral progress, has led many internalists vehemently to oppose 'generalism' – general, mediating moral principles or theory – in favour of 'particularism': just individual moral beliefs, X, Y, Z, motivating piecemeal.

What gives the darker counter-examples bite, is that it is plausible that certain very evil people in the world, are, and have been, some admixture of these three cases; (and that some very good people have been dissociated, generalist moralists). Internalism would be refuted if we established even the conceptual possibility of such cases, but an existence disproof is that much more compelling. There *are* evil people of differing kinds, some of whom *appear* on any independently justifiable grounds, to have normal enough abilities to discriminate cases as right or

wrong, and thus they appear to refute the following, claimed-as-conceptual entailment of the internalist:

> If Moral Belief X then *in-virtue*-of-that-belief, there will be a disposition to act X-morally-aptly.

The only way a conditional can be false, is of course for its antecedent to be true and its consequent false. So in response to the cases above, the following three strategies exhaust those available to the internalist.

1. Deny these cases of evil, etc., describe a situation where the antecedent holds true. So claim that the high-minded, or wicked, etc. person, despite appearances, doesn't *really* have the moral belief that X.
2. Deny that these cases describe a situation whereby the consequent is false. So claim that the high-minded, or wicked, etc. person, despite appearances, does *really* have *in-virtue*-of-that-belief, a disposition to act X-morally-aptly.
3. Get in place some exclusion clause as to the applicability of this conditional. So, 'If . . . then . . . *Except When* . . .'. E.g.:

i) . . . *Except When* there is a *weakness of the will*.[5]
ii) . . . *Except When* the person in question *isn't good*.[6]
iii) . . . *Except When* there *isn't*:
clear perception of the moral character of a situation gives, we have said, sufficient reason for action – which is not yet to say that action will ensue.[7]
iv) . . *Except When* the person in question would otherwise be *unintelligible* to us:
We must be able to elaborate the motivation of the wicked person in a way that would make his choice intelligible.[8]
v) . . . *Except When* the person is *irrational*, or
. . . practically irrational.

Now one point which is not typically at the forefront of these discussions but ought to be, is a little homily about intellectual

[5] See for example the Smith quote cited by fn. 3 above, and generally in the internalist literature.

[6] See for example the Smith quote cited by fn. 3 above, and generally in the internalist literature.

[7] Mark Platts, *Ways of Meaning* (London: Routledge and Keegan Paul, 1979), p. 261; and *passim* in the Oxford Wittgensteinian-Aristotelian literature.

[8] David McNaughton, *Moral Vision* (Oxford: Blackwell, 1988), p. 141; and generally in the internalist literature.

honesty. Any of the above may be acceptable counters, but there had better be reasons for embracing them independent of the desire to save internalism as a theory. In the light of this homily, I will not even begin to evaluate counters 3 i) to iv) for fear my problems with them would be *unintelligible* to my internalist opponents. Although what we will have to say about 3 v) probably does generalise to these other exclusion clauses, the reader is invited to assess them here as they stand. Rather, in the next section we will consider examples of the combined employment of strategies 1) and 2). In the section after this, we will draw out a picture of the philosophy of mind internalism is committed to as a result of employing these strategies, with a corresponding picture of how externalism opposes them in this. Finally we will turn to stratagem 3 v). Stratagem 3 v) is another way of disambiguating internalism as a thesis about *rationality*, and will be very important to the remainder of the paper.

IV

In the light of our homily above, what are we to make of the following example of a common internalist deployment of stratagem 1, here given by McDowell:

> might not another person have the same conception of the circumstances but see no reason to act as the virtuous person does? . . . We can evade this argument by denying its premise: that is, by taking a special view of the virtuous person's conception of the circumstances according to which it cannot be shared by someone who sees no reason to act as the virtuous person does.[9]

I take it that this is a straightforward No-true-Scotsman move: 'If he is a Scotsman, then it is conceptually necessary that he wears a kilt'. Objection – 'But yonder is a Scotsman in trews'. Response – 'Ah, but no true Scotsman would fail to wear a kilt'. (After all, that would be *inconceivable* . . .).

Is there any way to justify such stratagems? The majority of internalists who attempt to do so (eg., Dancy, Smith, McNaughton), appeal to a distinction, after Hare, between *Really* making a moral judgement, and making a moral judgement in

[9] 'Are Moral Requirements Hypothetical Imperatives?' *Procedings of the Aristotelian Society Supplementary* Vol. LII, 1978, p. 16.

'scare-quotes'. *Virtuous* people make *Moral* judgements, which they act on:

> If murder is *Judged* to be *Wrong*, then the virtuous person won't murder.

But amoralists, wicked and acidic people, only make 'moral' judgements:

> If murder is 'judged' to be 'wrong', the amoralist may murder anyway.

Unfortunately, since the criterion for *Really* making moral judgements – that is, making *Moral* judgements – is being virtuous (= *being motivated to act* on those judgements), this distinction seems merely to reify in jargon the no-true-Scotsman move.

An instructive variant on this attempt at justification is the intuition that if a subject 'knowingly' acts wrongly, this must be because she apprehends, say, that the action is wrong only in as much as it falls under a rule she accepts in general, not as *this action itself* being Really Wrong. Of course this again serves to beg the question, as the criterion for 'this action itself' being regarded as Really Wrong, will be the agent not being motivated to act thus; but, without retreating on this point, there are two other objections anyway. There are culpable sins of omission as well as commission, with this distinction itself usually being rather arbitrary. We are considering here a case of perspicuous epistemic access to a possessed moral directive. Whether we ignore the directive immediately or farm out the work by wilfully failing to subsume these individual cases under these general rules, or wilfully redirecting attention from a basic moral inference, the same thing is being done from the same base motives. Further, since culpable sophistry, intellectual dishonesty and so on are still *moral* failings direct, (quite apart from when, as considered here, in the service of other immoral ends), 'discovering' that our erstwhile moral failing was a case of such dishonesty would merely recapitulate the problem this manouvre was introduced to solve.

Note that the general approach of the internalist here is remarkably similar to a strategy exemplified in judicial uses of psychiatry, and polemicised as such by radical anti-psychiatrists like Thomas Szasz. We are presented with a real life amoralist, who has committed a gruesome series of murders and appears on independent grounds, well able to see that they are wrong. The

defence mounts a plea that he wasn't *really* able to see his deeds as *Wrong*, only 'wrong' in some cognitively more impoverished sense, as the defendant is a *Psychopath*. When the criterion for psychopathy is critically scrutinised and de-jargonised however, it appears to come down to little more than: 'person capable of monstrous, morally and socially contemptible acts, unintelligible by decent folk'. Lay jurors who reason similarly but without the permissive sanction of 'expert' medico-legal testimony, will be castigated for committing a *res ipsa loquitur* fallacy: 'the thing [monstrous act] speaks for itself'.

What is not so often noticed, is that the circularity of these moves is inevitable – is a necessary part of the internalist's theory as such. For internalists are making an internal, conceptual, 'criterial' connection between the relata of their conditional, above. The antecedent is some kind of *moral mental state*: a belief proper, 'conception of the circumstances', judgement, sentiment, perception, etc.. The consequent is some kind of action, motivation to act, *disposition to behaviour*, etc. In making a *conceptual connection* between mind and behaviour, internalism *just is* some kind of softer or harder logical behaviourism, at least as this applies to the special case of ethics.

That is, the price paid for a licence to print exclusion clauses, etc., *ex post facto* is not simply that the defence of this conditional will be ad hoc, but that it will succeed too well. The open-ended preparedness to always 'find out' that the consequent must have been true when the antecedent was – that there must have been a disposition to behaviour, however slight, given a moral belief – will inevitably make at least the consequent analytically necessary for the antecedent. (In fact, as defended in *use*, the internalist conditional quickly becomes an analytic biconditional, but establishing this is not necessary for my further arguments).

We are now in a position to see objections to internalism as being of a piece with realist objections to behaviourism – for instance, the *perfect actor* argument.[10] One imagines an actor, perfect at representing a raft of behaviours which the behaviourist holds to be conceptually sufficient for a mental state, yet who lacks that state. Or one imagines a person who has such a state, yet who lacks any of the behaviour the behaviourist claims

[10] See for example Hilary Putnam: 'Brains and behaviour' [1965] in: *Challenges to Empiricism*, ed. Harold Morick (London: Methuen 1980). The arguments of the next section can also be assimilated to certain anti-behaviourist counters.

to be conceptually necessary for that state (in ethics: one of the stock externalist figures of evil).

Given the operationalist, verificationist, tendencies of behaviourism, (hence, I have argued, internalism), a very quick objection is to be expected here: that even reporting a belief is a kind of behaviour. Any claim to have positively established the existence of *this* belief in *this* person will be held to require some behaviour, since behaviour is the 'criterion for' such a belief. So, the internalist may deny the existence of a moral belief in one of our moral monsters until some verification of its existence is forthcoming, then to insist that these feeble acts – say the discriminations made in completing a psychological test battery, or the verbal responses to a controlled interview – are the (conceptually necessary) behavioural criteria for the belief's existence.

The motivation behind such moves is the tacit assumption that externalism has not succeeded until it has produced for us an existence disproof: *this* belief sans any disposition to behaviour. However, to require such disproofs is a classic question-begging manoeuvre deployed by anti-realists across many areas, and to interpret externalism as aspiring to offer such a disproof is to deploy a classic anti-realist straw man. If, in order to prevail over the instrumentalist, the scientific realist had to produce evidence for a given electron sans photographic plates, meter-readings, etc. (ie sans any evidence for that electron), then realism would succeed only if it failed. Likewise for states of mind and their heterogeneous behavioural indicators: the externalist aspires to establish the conceptual independence – and indeed priority – of belief over disposition to behaviour, not *this* belief in *this* person, sans any behavioural evidence for *it*.

So, the perfect actor argument is that there is no internal connection between moral belief and disposition to behaviour because it is arbitrary how little disposition to behaviour is made conceptually necessary for the presence of a moral belief. Indeed, no degree more motivation to action is conceptually necessary given a moral belief, than one chooses to make criterial for the existence of the moral belief itself. To reject the notion of an analytic connection between moral belief and action, is then, not to

> . . . suppose (with the externalist) that moral facts are essentially inert.[11]

[11] Dancy, *Moral Reasons*, p. 255.

Rather, externalism is only committed to this:

Moral Facts are not essentially 'ert'.

(To establish that Taxi drivers are not *essentially* racist, it is not necessary that even one should be essentially non-racist).

V

The forgoing connects with two interlinked points: *commensurability* and *variance*. In rebutting the verificationist counter to the 'perfect actor' objection we see a limiting case of the former; in making the 'ertness' point, we have a case of the latter.

Commensurability: Internalism cannot be the anodyne claim that a moral belief (say, that it is wrong to slaughter the innocent), must as a matter of conceptual necessity produce some strength of disposition to behaviour from the whole range available, but without any conceptual constraint on *what* that strength of disposition to behaviour is – from high positive, through zero, to high negative. We saw that the amoralist and the sufferer from *accidie* may be defined as precisely people who possesses the moral belief that, say, slaughter of the innocent is wrong, but a vanishingly small disposition to act on that belief. For moral beliefs to be internally *connected* with disposition to action it will surely be necessary for the perceived moral *gravity* of the belief to be *commensurate* with its strength and direction of disposition to action. So to have a tiny, fatuous, disposition to act in virtue of a belief about the greatest evil, known as such to be easily preventable, is surely no less a confirmation of the externalist case from evil for that.

Further, a merely very good but defeasible correlation between perceived gravity of belief and disposition to act, would be wholly compatible with an externalist account of the connection between these as being causally mediated, (*vide* our criticism of the connection argument). Presumably invoking the concept of an internal connection requires something stronger here. However, it is then hard to see how *any* countervailing beliefs would be powerful enough to allow for the existence of certain kinds of savage evil.

Variance: The externalist can maintain that for beliefs, as for so many other things in nature, the strength of disposition to behaviour consequent on possessing any token of a given belief-type will be variably distributed in the human population – say, as a

normal distribution about a population mean. Such variance could account for the *oddness* of our extreme cases from evil and good, as well as their possibility, whether in the population of moral agents or in one individual's mind across time.

The possibility of such variance in *disposition* to behaviour, the externalist can accept in addition to accepting, in common with the internalist, a further source of variance in the actual behaviour of persons possessing the belief X, consequent on them possessing other beliefs which augment or countervail X's disposition. If we were to conceive of things in these terms, then in the light of the commensurability point, we would be likely to need to appeal to at least one other source of variance in addition to the latter, to account for actual behaviour that is grotesquely evil. (I make this point for those still in the grip of a picture whereby there are behavioural dispositions augmenting and countervailing one another, though in the next section I shall suggest we need something of a reconception of the issues here).

For a given moral belief-type, X, it is then a matter for empirical psychology to discover whether the population distribution of X includes any cases, however unusual, in which a token of X does not carry a disposition to action, or carries an inverted direction of disposition to action.

This claim is likely to invite the counter that it is somehow 'conceptually confused' to suppose that any *empirical* psychological discovery could establish the falsehood of internalism. Since the most conclusive way of establishing the conceptual possibility of a state of affairs is to show it is possible to subject it to empirical test, let us give an example of what such a test would be – of one way it would be for internalism to be false.

Non Associative Externalism: We establish a criterion on agreed independently motivated grounds for a subject to be in possession of a moral belief, X, Y, Z. 'Independently motivated' means to preclude any of the question-begging moves we have seen, whereby possession of a moral belief is denied unless evidence for its motivational force is forthcoming. In simplified (e.g. laboratory) cases, this would be easy. One's criterion for one's subjects being in possession of a moral belief, X, about, say, a hypothetical 'stimulus-person', S, is merely that one *tells them* S has (done) X. More naturalistic investigations would differ in degree of complexity and investigative ingenuity alone.

Then we either manipulate, or select on the basis of the world's manipulations, the presence of different combinations of

these beliefs in our groups of subjects, and we measure these subjects' behavioural responses for each combination of beliefs. In so doing we measure the behavioural effects of each belief when combined with various sets of other beliefs, and, for each such combination, when absent from an otherwise identical control set.

Now will each token of the same belief type be found to *add* some constant disposition (positive or negative) to whichever different set of beliefs it is combined with? (A linear, additive, *associative* model). Or will the difference the same belief, X, makes when added to some sets be to effect a magnification, to other sets a division, to other sets a reversal, and to other sets no difference at all, in the behaviour that would have been present without it? (A non-linear, interactive, *anti-associative* model).

Suppose we are measuring in simulated judges the inclination to punish a wrongdoer. Let our moral belief be that he has been shamefully and obnoxiously drunk. I suggest that in the presence of a belief set which includes the belief that the subject has abandoned his wife and children, our drunkenness belief will make one contribution to punitive behaviour; in the presence of a belief set that includes the belief that the subject has been abandoned himself it will make a quite other contribution. For in each case the shared belief will *interact* with the rest of the belief set differently because of the beliefs that are not shared.

In any event, this predicted effect will differ measurably and testably from the associative model which would have it that in the one case the constant disposition to punish is overwhelmed by the mere *addition* of a countervailing pity, (to which it nevertheless subtracts its disposition), and in the other case a contempt for the man's act of desertion is augmented by that same constant disposition.

Let us stick with beliefs about drunkenness, though for a different example. The attribution of drunkenness to an agent may non-additively *magnify* our punitive dispositions in some circumstances (accidental death by driving, say), yet its presence may *diminish* our punitive dispositions in other cases (reducing the moral condemnation associated with homicide from murder to manslaughter). One can of course insist that these beliefs about homicide, drunkenness, etc., are not finished moral beliefs, and that when finished (i.e., combined), we do not get two tokens of the same belief type. That, though, looks perilously question-begging – it looks to

make it impossible for two moral beliefs ever to be tokens of the same type, save when we have global cognitive identity – and in preventing the dispositional properties of the whole from being predictable from the sum of their parts, it would anyway serve our anti-associationist purposes just as well; terminology of the 'parts' notwithstanding.

The claim being made here is that the contribution any given token of a belief type, X, will make to eventual behaviour, will vary in strength, and even direction, according to the other beliefs in the belief set because it will interact with these other beliefs. *No given moral belief will have a constant strength or direction of motivation to action associated with it a priori.* Strength (from zero upwards), and direction of motivation, will be radically dependent on the other beliefs it enters into combination with in that agents belief set, (better: that agent's *mind*). Behaviour is not in any sense an additive resultant of all the prior behavioural dispositions, positive and negative, in a person's belief set. There are no such prior dispositions.

The belief that it is wrong to steal, will not, for instance, simply and constantly dispose one not to steal to such and such a degree, with merely the possibility of countervailing constant dispositions from other beliefs preventing this. It is not even true a priori, that of two agents with otherwise identical belief sets, the one with a strength of conviction three times more powerful than the other that stealing is wrong will be less likely to steal. In the presence of the shared belief that an iniquitous society, say, sanctions grotesque and hypocritical violations of that moral precept, and shared beliefs about the absence of other forms of redress, etc., it may make *more likely* that the stronger anti-theft moralist steals, as compared with the agent who lacks the warping produced by that strength of moral indignation to begin with. It is highly dubious to respond to such cases by attributing to the subject a qualified or unfinished belief ('wrong to steal, *except when*'), as for any mature moral agent these qualifications would surely be indefinitely long for each belief.

Similarly, an accurate moral belief in the mind of an evil person need not carry the same positive disposition to action as it does in the mind of a good person, with the evil agent's eventual behaviour merely the result of the subtraction of countervailing (inaccurate) moral beliefs. It is not conceptually confused to suppose that a moral cognition, known to be

accurate, should nevertheless sometimes count for nothing even as *disposition* to behaviour in the mind of an evil person, it is quite possible it should even make *more* likely that evil behaviour results.

It is hard to see what 'conceptual' objections to these claims amount to if not to an inability to see that it is at least *conceivable* that atomistic associationism should be false. The suggestion here is that we are not simply moral adding machines. Before we can know anything of how, or whether, the possession of a moral belief will motivate, we must know a good deal about the *mind* it operates within.

VI

I take it that we have, with the above, established at least the conceptual possibility of an agent having a moral belief, X, yet not in-virtue-of-that-belief having the disposition to act X-morally-aptly. So unless stratagem (3) works to save the internalist conditional, the variety of internalism based on disambiguation (a) of 'reasons for action', will have been refuted. And we have already seen how internalism, in claiming that moral beliefs are 'reasons for action', cannot mean purely *moral* reasons, (reasons (c)) for this, we saw in considering the argument from oddness, is something the externalist is strongly committed to, while arguably the *internalist* is not.

Thus, in considering stratagem 3(v) (the 'except when irrational' exclusion), we just are considering the only disambiguation of internalism left: that moral beliefs are normative *epistemological* reasons for action – reasons (b). Meaning that it is conceptually impossible for someone who is *rational* not to be motivated by her moral beliefs.

Notice what each version of internalism is struggling with here: accounting for *evil*. Internalism$_{(a)}$ held that to know the Good and not be disposed to act on it was inconceivable. One can of course, fail to know the Good – internalism has no problems possessing an account of *Badness*. Badness is ignorance of the good. Internalism$_{(a)}$ has, however, real problems with the radically evil person – one who is not ignorant of the good, but just does not act on it. If the criterion for possession of a moral belief is made into the possession of the requisite disposition to behaviour, no sense can be made of the sceptical worry expressed thus:

'I know it's wrong to do this, but why should I care about being moral?'

Internalism$_{(b)}$ can make sense of this worry, but only in the *irrational*. For Internalism$_{(b)}$ the assumption seems to be that if we can convict the evildoer of irrationality, we have a knockout blow, a headlock to drag him back into morality. But of course, he can always reply: 'So I'm irrational – why should I care about being rational', and the sceptical worry merely resurfaces. One might want to avail oneself of another *except when* . . . stratagem here in turn: *except when* of poor character, or weak of will, or incontinent, or what have you. But to avoid a regress, sooner or later a species of internalism must be advanced without any get-out clause – and suppose for argument's sake this is rationality internalism: it really is conceptually impossible *full stop* for anyone capable of discovering they are irrational, not thereby to be motivated to cease being so.

Now as a minor point: any such claim does seem to amount to a rather arbitrary, even arrogant, extension of the powers of the philosopher. Who is to tell the psychiatrist she is *conceptually confused* to have written on patient X's notes that 'he knows he is being irrational and doesn't care'? Is one to tell her, without any clinical contact with X, that it should be 'knows' rather than *Knows*?

More significantly, a non-regressive internalism$_{(b)}$ looks to have to rule out the *rational Nazi* a priori – at least for a bona fide, mass murdering Nazi. Granted that such a figure is *evil*, he cannot then be *rational*. Well, I am severely sceptical as to whether there can be any independent motivation for claiming this, but grant it anyway. Still the point at issue can be left at this: is the Nazi's alleged irrationality *commensurate with* his avowed evil?

In line with the critical approaches of the last section, note that we don't have to establish that the Nazi may be wholly, or even largely rational (though I happen to suppose that the latter is possible and the former is conceptually impossible for anyone). It would be open to us to accept that the Nazis were irrational, but to point out that the guards at Auschwitz could be very irrational, yet still there will be manifestly greater irrationality in any asylum, without one whit of the *evil*. It is not legitimate to respond here that the Nazis were really, *Really*, more irrational than the benign inmates of these asylums, for the only ground for saying that will be to redefine as especially

'irrational' their (non-evident) irrationality because it is especially, evidently, *blatantly* evil. Put another way, one may not partition a special sub-class of irrationality – purely 'moral' irrationality – which is distinct from the everyday varieties found in asylums and on omnibuses, for that is to make the slide into 'reasons (c)' once again, and beg the question.

Nor can the internalist simply rest with the claim that the Nazis necessarily were irrational to some degree. For one thing, everyone is. Internalism$_{(b)}$ is claiming that immorality is internally *connected* with irrationality. Establishing (if it can) that the Nazis were necessarily irrational to at least this small degree, given that they were immoral to this huge degree, does not begin to do this. Besides, everyone admits these harmless schizophrenics are irrational to a huge degree, and not evidently evil at all. If evil is to be internally connected with an error of rationality, then it has to be shown that the degree of that error, specified on its own, independently motivated terms, and the degree of evil, specified on its independent terms, must covary together and *cannot diverge*.

One might think this too strong a requirement to saddle internalism with, but note, firstly, that a merely good correlation is wholly compatible with any externalism that (in my opinion, unwisely), chooses to derive its ethics from its epistemology. As we saw in our criticism of internalism$_{(a)}$'s use of the connection argument, externalism can predict a close but defeasible correlation – here between the presence of irrationality and the presence of evil – one mediated by causation. Internalism must be arguing for something stronger than a defeasible correlation.

Nor, secondly, could internalism remain a serious position if it repaired to a 'two-stage' theory, whereby only the bare connection between being immoral and being less than perfectly rational were held to be a priori, while conceding that the connection between the degree of that error of rationality and magnitude of evil was to be left a posteriori. (Or put another way, a theory only that one has an a priori reason to act on one's moral beliefs to the extent that one is perfectly rational, yet with the strength of that motivation – at least in the less than perfectly rational – left a posteriori). Since we are none of us wholly rational, we will have here a terminological sop to internalism tacked on to a theory of moral motivation that is a posteriori and external in all but name.

Although, as indicated, it would be open to externalists to derive their ethics from their epistemology, clearly my judgement from the above is that certain kinds of immorality are not any kind of error of rationality, they represent the *success*, the *triumph*, of evil.

This point generalises importantly beyond rationality internalism alone. Whether one stops with evil as ignorance, or irrationality, or imperfect ('cloudy') moral 'vision', or poorly formed *character*, or Aristotelian *incontinence*, or *weakness of will*, or some unspecified cocktail of such things; in the limit, there will always be something missing, because each of these are *performance errors* of a basic human cognitive-conative *competence* that is essentially morally good. (Employing here, the competence/performance distinction common in the psychological sciences after Chomsky).

For internalism$_{(a)}$ the evil person can only be ignorant, for Internalism$_{(b)}$ the evil person can rather be irrational. Mix and match as many moral performance errors as you like, still, at the end: Sad or Mad, but never just plain Bad. That is, internalism can have an account of the person who *does* bad (through ignorance, irrationality, or some other subversion of their pure moral competence), but not an account of the person who *is* bad – is evil. However much qualified by exclusion clauses, Internalism can have an account of evil only as a *performance error* of some sort – a *performance*, but not a *competence* theory of evil. I see this as a core commitment of internalism, and as one which will inevitably deny us adequate resources to evaluate the darker side of humanity overall.

VII

Premise One, which we began with, is then to be rejected on the first two disambiguations of 'reasons' considered. Moral 'beliefs' alone, do not as a matter of *conceptual* necessity, provide either psychological or epistemological reasons for action. On the third such disambiguation of 'reasons' – the purely ethical – Premise One is true, but does not describe any species of moral *internalism*. If the versions of Premise One we have looked at exhausted the options, then internalism would stand refuted.

Of course there are responses open to the internalist here. Other disambiguations of Premise One will no doubt be available, and, (which may come to the same thing), other exclusion

clauses for the disambiguations considered. What can be said in advance, is that other putative disambiguations or exclusions must ensure they do not fall foul of the general cautions we have noted. In particular, that any defence of an internalist conditional against refutation, must be motivated independently of the need to save internalism as a theory – *i.e.*, must not involve a parade of no-true-Scotsman moves. Also, that any variety of internalism must confront square on the apparent entailment of internalism just noted: that it makes the nature of human moral cognitive-conative *competence* essentially good, whatever 'errors' the *performance* may bring. For considered as a putative species of realism, cognitivist internalism appears unlikely to offer us a robust enough picture of our tendencies to act (sometimes *intentionally*, sometimes *culpably*) at odds with what we know to be right. And here I would ask my reader simply to re-couch her favoured version of the internalist conditional into its contrapositive and contemplate the result.

In considering this contraposition, we see Internalism for what, at the most abstract level it is – a species of ethical determinism: that, necessarily, we are determined to moral action by what we take to be the right. (Kant: we are not volunteers in the army of duty[12]). Disputes between cognitivist and non-cognitivist internalists can then be seen as disputes about the psychological specifics of this determination: beliefs alone, or beliefs plus desires. One can then understand the tendency of cognitivist internalist to assume that their (cognitivist) externalist opponents must share an adherence to a version of Premise Two with their non-cognitivist internalist opponents. Externalism claims we may have a moral belief yet not thereby be disposed to morally relevant action on it. The motivation behind this is to leave space for morally culpable and commendable responses to our moral beliefs, insisting that we have some choice in how we act on our moral beliefs. That means we can, in some sense, *choose* to be moral or immoral.

It is then assumed that the externalist's choice must consist in some kind of *desire* to be moral and that moral imperatives become *hypothetical* on possession of that desire. The internalist plausibly insists this further step is unwarranted and regressive. Why not just say that without the disposition to action there isn't the requisite belief? Surely no net gain is to be made in the

[12] See Phillipa Foot, *Virtues and Vices* (Oxford: Blackwell 1978), p. 170.

direction of moral freedom, culpability, and a positive (competence) theory of evil, by insisting that we aren't always determined to act on what we *believe* to be right, because a determinant from *desire* may also be needed? I mean, this latter might be psychologically *true* as a detail, but it is not any nearer to a robust moral freedom. To say freedom – here moral – consists in an agent's being 'free' (from constraint) in exercising her (determined) desires is Hobbes' and Hume's soft determinism. Some externalists may choose to take that path, but I do not; moral freedom, real choice, is not the exercise of an unfrustrated (determined) desire, any more than it is the exercise of a determined true belief. It is the *self determination of action* in response to one's moral beliefs.

The error here was to suppose that if two agents differ in motivation to X-morally-relevant action yet do not differ in the moral belief that-X, there (logically) must be another 'ingredient' – whether desire or other – which makes the difference. Unless 'desire' here is merely a harmless synonym for choice, this is the associative psychological framework holding us in its baleful grip once again. Persons choose actions (in response to their beliefs), beliefs themselves don't, desires don't, nothing else does. Persons may be *composed* of nothing but mental items like beliefs, desires, (or other, scientific items), but they are not identical to these items; the properties of the whole are different to the properties of its parts. Our minds are not bundles of mental items, each adding its little push or pull. When the same moral belief produces different motivation to action in different people this is because they have (morally) different minds. The choice of how to respond to a moral belief is a self-determination of the person, their mind, their conscience.

I offer these last remarks rather tentatively, to point out, firstly, one way an attribution to the externalist of an adherence to Premise Two, above, can be resisted. And secondly, that internalists' problems with the positive externalist view are going to amount to very general metaphysical objections to the notion of real moral freedom: the notion that there could be external standards of rightness which people can conceivably know of yet *choose* to live up to or not.

Whatever responses are available to them, I hope that with these criticisms we have at least established that internalists have consistently underestimated the work they have to do to defend their thesis, and have consistently overestimated how strong any

externalist position would have to be which opposed it. The aim of this paper has been to show that we have good grounds for resisting internalism on quite weakly realist assumptions – whether in the philosophy of mind, general metaphysics, or ethics per se.[13]

Psychology Department,
University of Luton,
Luton, Bedfordshire, LU1 3JU

[13] A version of this paper was read at the St. Andrews 'Ethics and Practical Reasoning Conference' in March 1995. My thanks go out to the conference organisers and fellow participants; also to Steve Wilkinson, Jonathan Dancy, Richard Gaskin and Michael Morris for comments on an earlier, written, version.