Philosophy and Phenomenological Research
Vol. LV, No. 1, March 1995

Internal Reasons

MICHAEL SMITH
Australian National University

Introduction

According to one popular version of the dispositional theory of value, the version I favour, there is an analytic connection between the desirability of an agent's acting in a certain way in certain circumstances and her having a desire to act in that way in those circumstances if she were fully rational (Rawls 1971: Chapter 7; Brandt 1979: Chapter 1; Smith 1989, 1992, 1994).[1] If claims about what we have reason to do are equivalent to, or are in some way entailed by, claims about what it is desirable for us to do—if our reasons follow in the wake of our values—then it follows that there is a plausible analytic connection between what we have reason to do in certain circumstances and what we would desire to do in those circumstances if we were fully rational.

The idea that there is such an analytic connection will hardly come as news. It amounts to no more and no less than an endorsement of the claim that all reasons are 'internal', as opposed to 'external', to use Bernard Williams's terms (Williams 1980). Or, to put things in the way Christine Korsgaard favours, it amounts to an endorsement of the 'internalism requirement' on reasons (Korsgaard 1986). But how exactly is the internalism requirement to be understood? What does it tell us about the nature of reasons? And where-in lies its appeal? My aim in this paper is to answer these questions.

The paper divides into three main sections. In the first I distinguish between two different models of the internalism requirement—the 'advice' model and the 'example' model—and I say why the requirement should be understood in terms of the advice model. In the second and longest section I spell out the requirement in some detail and I explain why, contrary to

[1] Adherents of other versions of the dispositional theory may agree that desirability is a feature that elicits an appropriate response in subjects under conditions of full rationality, but disagree about whether that response is desire (Johnston 1989 appears to take this view), or they may instead agree that desirability is a feature that elicits desire in agents under the appropriate conditions, but disagree about whether those are conditions of full rationality (Lewis 1989 appears to take this view).

Bernard Williams, it is not especially allied to a relativistic conception of reasons—indeed I say why those of us who embrace the requirement should endorse a non-relative conception. And in the third section I use the advice model, understood in the way explained in the second section, to explain the appeal of the internalism requirement. As we will see, the internalism requirement helps us solve an otherwise troubling problem about the effectiveness of deliberation.

1. The advice model versus the example model

The internalism requirement tells us that the desirability of an agent's ϕ-ing in certain circumstances C depends on whether she would desire that she ϕs in C if she were fully rational. This idea can be made more precise as follows.

We are to imagine two possible worlds: the *evaluated* world in which we find the agent in the circumstances she faces, and the *evaluating* world in which we find the agent's fully rational self. In these terms, the internalism requirement tells us that the desirability of the agent's ϕ-ing in the evaluated world depends on whether her fully rational self in the evaluating world would desire that she ϕs in the evaluated world. Note what I have just said, for the precise formulation is important. The idea is that we are to imagine the agent's fully rational self in the evaluating world looking across at herself in the evaluated world (so to speak) and forming a desire about what her less than fully rational self is to do in the circumstances she faces in that evaluated world. We might imagine that the self in the evaluating world is giving the self in the evaluated world advice about what to do. Accordingly, this is what I call the 'advice' model of the requirement.

The advice model of the requirement contrasts with the example model. On this alternative way of thinking about the requirement, the idea is that the desirability of an agent's ϕ-ing in the evaluated world depends on whether her fully rational self in the evaluating world would desire to ϕ *in the evaluating world*. We are not to suppose that the agent's fully rational self is giving advice to herself in the evaluated world, but rather that the agent's fully rational self is setting up her own behaviour in her own world, the evaluating world, as an example to be followed by the self in the evaluated world. The issue of interpretation, then, turns on whether the internalism requirement tells us that in acting on reasons we follow the *advice*, or the *example*, of our fully rational selves.

I said that the details of the formulation are important, and the reason why is because the details are something about which those who accept the requirement may yet disagree. Consider, for example, Christine Korsgaard's own official formulation of the requirement. According to Korsgaard the internalism requirement is the claim that the considerations that constitute reasons must 'succeed in *motivating* us insofar as we are rational' (1986: 15,

emphasis is mine). But on the plausible assumption that a fully rational agent's desires will only succeed in motivating *her* if they are desires that concern the circumstances in which *she* finds *herself*, the idea, in our terms, must be that a consideration constitutes a reason in the evaluated world just in case, in the evaluating world, the agent's fully rational self would desire that she acts on that consideration *in the evaluating world.* Korsgaard thus seems to have in mind the example model of the internalism requirement, not the advice model.[2]

But the example model is plainly wrong. In order to see why consider the following case, a variation on an example of Gary Watson's (1975). Suppose I have just been defeated in a game of squash. The defeat has been so humiliating that, out of anger and frustration, I am consumed with a desire to smash my opponent in the face with my racket. But if I were fully rational, we will suppose, I wouldn't have any such desire at all. My desire to smash him in the face is wholly and solely the product of anger and frustration, something we can rightly imagine away when we imagine me in my cool and calm fully rational state. The consideration that would motivate me if I were fully rational is rather that I could show good sportsmanship by striding right over and shaking my opponent by the hand. In that case, does it follow that what I have reason to do *in my uncalm and uncool state* is stride right over and shake him by the hand?

In essence, this is what Korsgaard's formulation of the internalism requirement tells us, for she supposes that a consideration constitutes a reason just in case it would motivate the fully rational person, and this is what my fully rational self would be motivated to do. And yet this is surely quite wrong. Striding right over and shaking my opponent by the hand might be the last thing I have reason to do, especially if being in such close proximity to him, given my anger and frustration, is the sort of thing that would cause me to smash him in the face. Rather, we might plausibly suppose, what I have reason to do in my uncalm and uncool state is to smile politely and leave the scene as soon as possible. For this is something that I can get myself to do and it will allow me to control my feelings. Moreover—and importantly for the advice model—*this is exactly what my fully rational self would want my less than fully rational self to do in the circumstances that my less than fully rational self finds himself.* But, to repeat, it is not something I would be motivated to do if I were fully rational because it is not something that I would have any *need* to be motivated to do if I were fully rational.

---

2 Rawls (1971) and Brandt (1979) seem to have had in mind the example model of the internalism requirement as well. Contrast Peter Railton's account of a person's own good (1986) which is formulated in terms in terms of the advice model precisely to avoid problems like those I go on to describe in the text. For criticisms of Rawls's and Brandt's 'example' versions of the internalism requirement see Shope (1978) and Pettit and Smith (1993). (Here I am grateful to Stephen Darwall.)

The example model of the internalism requirement thus gives us the wrong answer in cases in which what we have reason to do is in part determined by the fact that we are irrational. For what an agent's fully rational self is motivated to do will depend on the circumstances in which she finds herself, and, by definition, these circumstances will never include her own irrationality. It therefore seems to me that we should reject the example model of the internalism requirement in favour of the advice model. What we have reason to do in the circumstances in which we find ourselves is fixed by the advice our fully rational selves would give us about what to do in these circumstances that we face.

2. The internalism requirement and the idea of being fully rational

The internalism requirement tells us that it is desirable for an agent to ϕ in certain circumstances C, and so she has a reason to ϕ in C, if and only if, if she were fully rational, she would desire that she ϕs in C. The content of our reasons is thus fixed by the advice we would give ourselves if we were fully rational. However, note that I haven't yet said anything about what being 'fully rational' means, and that we must do so if we are to understand what the internalism requirement tells us, substantively, about the reasons we have.

In his own similar analysis of internal reasons Bernard Williams suggests, in effect, that to be fully rational in the practical sphere an agent must satisfy the following three conditions:

(i) the agent must have no false beliefs

(ii) the agent must have all relevant true beliefs

(iii) the agent must deliberate correctly

His reason for insisting on the first two conditions is straightforward enough.

If our desire to do something is wholly dependent on false beliefs, then we ordinarily suppose that it isn't really desirable to do that thing. Suppose, for example, I desire to drink from a particular glass, but that my desire to do so depends on my belief that the glass contains gin and tonic when in fact it contains gin and petrol. Then we would ordinarily say that though I might think that it is desirable to drink from the glass, it isn't really desirable to do so. Why not? Because I would not desire that I do so if I were fully rational: that is, if, *inter alia*, I had no false beliefs—thus condition (i).

Similarly, in the case of condition (ii), if we fail to desire something, and if our failure to do so is wholly dependent on our failure to believe something that is true, then we ordinarily suppose that that thing may yet be desirable. Suppose, for example, that I do not desire to drink from a particular glass,

but that my failure to do so is to be explained by the fact that I am ignorant of the contents of the glass. In fact it contains the most delicious drink imaginable. Then we would ordinarily say that despite the fact that I do not desire to drink from the glass, doing so may yet be desirable. Why? Because I may well desire that I do so if I were fully rational: that is, if, *inter alia*, I had all relevant true beliefs.

But what about condition (iii)? Williams's idea here is that even if we fail to desire that we ϕ, ϕ-ing may still be desirable because we would desire that we ϕ if our other beliefs and desires interacted in the ways appropriate for the generation of new desires: that is, if we deliberated and did so correctly. For example, the means to an end is desirable, but we will in fact desire the means to our ends only if we reason in accordance with the means-ends principle, for only so does a desire for an end turn into a desire for the means.

Moreover, as Williams points out, means-ends reasoning is only one mode of rational deliberation among many. Another example is

> ...practical reasoning...leading to the conclusion that one has reason to ϕ because ϕ-ing would be the most convenient, economical, pleasant etc. way of satisfying some element in...[one's set of desires]...and this of course is controlled by other elements in...[one's set of desires]...if not necessarily in a very clear or determinate way....[And]...there are much wider possibilities for deliberation, such as: thinking how the satisfaction of elements in...[one's set of desires]...can be combined: e.g. by time-ordering; where there is some irresoluble conflict among the elements of...[one's set of desires]...considering which one attaches most weight to...; or, again, finding constitutive solutions, such as deciding what would make for an entertaining evening, granted that one wants entertainment. (1980: 104)

And he thinks that there are other, more radical, possibilities for deliberation as well.

> More subtly,...[an agent]...may think he has reason to promote some development because he has not exercised his imagination enough about what it would be like if it came about. In his unaided deliberative reason, or encouraged by the persuasions of others, he may come to have some more concrete sense of what would be involved, and lose his desire for it, just as positively, the imagination can create new possibilities and new desires. (1980: 104–5)

Thus, according to Williams, we must include the operation of the imagination in an account of what is involved in deliberating correctly as well.

Williams's conditions (i) through (iii) seem to me to constitute a fairly accurate spelling out of our idea of what it means to be practically rational. An agent who has defective beliefs or who deliberates badly is indeed the sort of agent we tend to think of as being practically irrational in some way. It seems to me that Williams's conditions do require supplementation and amendment, however. For one thing, I see no way in which the effects of anger and frustration could be precluded by conditions (i) through (iii) —unless some such constraint is supposed to be presupposed by condition (iii), the condition of correct deliberation. Yet, as we have seen, emotions can

cause us to desire to do what we have no reason to do (remember the effects of that humiliating defeat I suffered in squash). Here, then, there is need for supplementation. And for another—and this is the point on which I wish to focus—it seems to me that Williams omits from his discussion of condition (iii) an account of perhaps the most important form of deliberation. The omission is serious as it leads him to overstate the role of the imagination in deliberation. Here, then, as we will see, there is need for both supplementation and amendment.

Williams admits that deliberation can produce new and destroy old underived desires. As he puts it, an agent 'may think he has reason to promote some development because he has not exercised his imagination enough about what it would be like if it came about', just as, more 'positively, the imagination can create new possibilities and new desires'. When the imagination does create and destroy desires in these ways Williams tells us that we take its operations to be sanctioned by reason.

Williams is right, I think, that deliberation can both produce new and destroy old underived desires. But he is wrong that the only, or even the most important, way in which this happens is via the exercise of the imagination. By far the most important way in which we create new and destroy old underived desires when we deliberate is by trying to find out whether our desires are, as a whole, *systematically justifiable*. And, if this is right, then that in turn requires a significant qualification of Williams's claim that reason sanctions the operation of the imagination.

What do I mean when I say that we sometimes deliberate by trying to find out whether our desires, as a whole, are systematically justifiable? I mean just that we can try to decide whether or not some particular underived desire that we have or might have is a desire to do something that is itself nonderivatively desirable, and that we do this in a certain characteristic way: namely, by trying to integrate the object of that desire into a more *coherent* and *unified* desiderative profile and evaluative outlook. Rawls describes the basics of this procedure of systematic justification in his discussion of how we attempt to find a 'reflective equilibrium' among our specific and general evaluative beliefs (Rawls 1951; Daniels 1979). I will restrict myself to saying a little about the way in which achieving reflective equilibrium may also be a goal in the formation of underived desires.

Suppose we take a whole host of desires we have for specific and general things, desires which are not in fact derived from any desire we have for something more general. We can ask ourselves whether we wouldn't get a more systematically justifiable set of desires by adding to this whole host of specific and general desires another general desire, or a more general desire still, a desire that, in turn, justifies and explains the more specific desires that we have. And the answer might be that we would. If the new set of desires—the set we imagine ourselves having if we add a more general desire to the

more specific desires we in fact have—exhibits more in the way of coherence and unity, then we may properly think that the new imaginary set of desires is rationally preferable to the old. For the coherence and unity of a set of desires is a virtue, a virtue that in turn makes for the rationality of the set as a whole. This is because exhibiting coherence and unity is partially constitutive of having a systematically justified, and so rationally preferable, set of desires, just as exhibiting coherence and unity is partially constitutive of having a systematically justified, and so rationally preferable, set of beliefs.

The idea here is straightforwardly analogous to what Rawls has to say about the conditions under which we might come to think that we should acquire a new belief in a general principle given our stock of rather specific evaluative beliefs. The thought there is that we might find that our specific value judgements would be more satisfyingly justified and explained by seeing them as all falling under a more general principle. The imaginary set of beliefs we get by adding the belief in the more general principle may exhibit more in the way of coherence and unity than our current stock of beliefs. Likewise, the idea here is that our imaginary set of desires may exhibit more in the way of coherence and unity than our current set of desires.

If we do come to believe that our more specific desires are better justified, and so explained, in this way, then note that that belief may itself cause us to have a new, underived, desire for that more general thing. And, if it does, then it seems entirely right and proper to suppose that this new desire has been arrived at by a rational method. Indeed, the acquisition of the new more general desire will seem rationally required in exactly the same way that the acquisition of the new belief that the object of the desire is desirable will seem rationally required. In fact, if the internalism requirement is right, the acquisition of a new evaluative belief will be the cognitive counterpart of the acquisition of the new desire. For, according to the requirement, an evaluative belief is simply a belief about what would be desired if we were fully rational, and the new desire is acquired precisely because it is believed to be required for us to be more rational.

Moreover, if this is agreed, then note that we can not only explain how we might come to have new underived desires as the result of such reflection, but that we can also explain how we might come to lose old underived desires as well. For, given the goal of having a systematically justifiable set of desires, it may well turn out that, as the attempt at systematic justification proceeds, certain desires that seemed otherwise unassailable have to be given up. Perhaps because we can see no way of integrating those desires into the set as a whole they will come to seem *ad hoc* and so unjustifiable to us. Our belief that such desires are *ad hoc* may then cause us to lose them. And, if so, then it will seem sensible to describe this as a loss that is itself mandated by reason; as again straightforwardly analogous to the loss of an unjustifiable, because *ad hoc*, belief.

As this procedure of systematic justification continues we can therefore well imagine wholesale shifts in our desiderative profile. Systematic reasoning creates new underived desires and destroys old. Since each such change seems rationally required, the new desiderative profile will seem not just different from the old, but better; more rational. Indeed, it will seem better and more rational in exactly the same way, and for the same reasons, that our new corresponding evaluative beliefs will seem better and more rational than our old ones.

To a first approximation, then, this is what I mean by saying that we can create new and destroy old underived desires by trying to come up with a set of desires that is systematically justifiable. But even this first approximation is enough to see why Williams's claims about the role of the imagination in deliberation requires significant qualification. For true though it is that the imagination can produce new and destroy old underived desires via vivid presentations of the facts, its operations are not guaranteed to produce and destroy desires that would themselves be sanctioned in an attempt at systematic justification of the kind just described. In fact quite the opposite is the case. For the imagination is liable to all sorts of distorting influences, influences that it is the role of systematic reasoning to sort out. Consider an example. Vividly imagining what it would be like to kill someone, I might find myself thoroughly averse to the prospect no matter what the imagined outcome. But, for all that, I might well find that the desire to kill someone, given certain outcomes, is one element in a systematically justifiable set of desires. Merely imagining a killing, no matter what the imagined circumstances, may cause in me a thoroughgoing aversion, but it will not justify such an aversion if considerations of overall coherence and unity demand that I have a desire to kill in certain sorts of circumstances, and such considerations may themselves override the effects of the imagination and cause me to have the desire I am justified in having.[3] The role played by attempts at systematic justification is thus what is crucially required for an understanding of how deliberation creates new and destroys old underived desires, not the role played by the imagination.

Let's recap. According to the internalism requirement, the desirability of an agent's ϕ-ing in certain circumstances C is fixed by whether or not she would desire that she ϕs in C if she were fully rational. The aim in this section is to spell out the idea of being fully rational. Taking our lead from Bernard Williams the suggestion so far is that an agent is fully rational just in case she has no false beliefs and all relevant true beliefs, and just in case she deliberates correctly in the light of these beliefs, and an agent is in turn understood to have deliberated correctly just in case her underived desires are

---

[3] Mark Johnston (1989) pursues a similar line in his criticism of David Lewis's account of the role of imaginative acquaintance in valuing (1989).

systematically justifiable: that is, to a first approximation, just in case her underived desires form a maximally coherent and unified desire set. Do we need to say more? Indeed we do, something we see clearly once we focus on a consequence Williams wants us to draw from his own similar analysis of reasons.

According to Williams, the internalism requirement supports a *relative* conception of reasons. He puts the point this way.

> [T]he truth of the sentence...['A has a reason to ϕ']...implies, very roughly, that A has some motive which will be served or furthered by his ϕ-ing, and if this turns out not to be so the sentence is false: there is a condition relating to the agent's aims, and if this is not satisfied it is not true to say...that he has a reason to ϕ. (1980: 101)

And again later:

> Basically, and by definition,...[an analysis of reasons]...must display a relativity of ...[a]...reason statement to the agent's *subjective motivational set*...(1980: 102)

Now in fact it is initially quite difficult to see why Williams says any of this at all. For, as we have seen, what the internalism requirement suggests is that claims about an agent's reasons are claims about her *hypothetical* desires, not claims about her *actual* desires. The truth of the sentence 'A has a reason to ϕ' thus does not imply, not even 'very roughly', that A *has* some motive which will be served by his ϕ-ing; indeed A's *motives* are beside the point—that was the difference between the advice model and the example model. What the internalism requirement implies is rather that A has a reason to ϕ in certain circumstances C just in case he *would* desire that he ϕs in those circumstances if he were fully rational.

Williams might concede this. But, he might say, it doesn't show that he is wrong when he says that the requirement supports the relativity of an agent's reasons to her actual desires, it simply shows that the relativity of reasons requires more careful formulation. The crucial point, he might insist, is that the desires an agent would have if she were fully rational are themselves simply functions from her actual desires, where the relevant functions are those described in conditions (i) through (iii). An agent's reasons are thus relative to her actual desires, he might say, because under conditions of full rationality agents would all have different desires about what is to be done in the various circumstances they might face. Even if it is rational for each of us to change our actual desires by trying to come up with a set of desires that can be systematically justified, in the manner captured by conditions (i) through (iii), such changes will always fall short of making us have the same desires as our fellows; they will always reflect the antecedent fact that we have the actual desires that we have. The content of the maximally coherent and

unified desire set any particular agent could have will always reflect the content of that agent's actual desires.

As I see it, this is what Williams has in mind when he says that our reasons are all relative.[4] It explains why he rightly insists that he is defending a 'Humean' conception of reasons (1980: 102). For his conception of reasons, like Hume's own, is predicated on skepticism about the scope for reasoned change in our desires (Korsgaard 1986); predicated on denying that, through a process of rational deliberation—through attempting to give a systematic justification of our desires, for example—we could ever come to discover reasons that we all share. For what we have reason to do is given by the content of the desires we would have if we were fully rational, and these may differ in content from agent to agent.

Williams claims to derive this relative conception of reasons from the internalism requirement. But as a *derivation* this is hardly compelling. It goes through only if we assume that it is no part of our task, in trying to come up with a systematically justifiable set of desires, to come up with the same set of desires as our fellow rational creatures would come up with if they set themselves the same task. And this suggests, in turn, that there are therefore two quite distinct conceptions of *internal* reasons. There is a relativistic, Humean, conception of internal reasons—the conception embraced by Williams—and there is also a non-relativistic, anti-Humean or Kantian conception according to which, if we were to engage in a process of systematically justifying our desires we would all eventually reason ourselves towards the same conclusions as regards what is to be done. That is, according to the opposing conception, all possible rational creatures would desire alike as regards what is to be done in the various circumstances they might face because this is, *inter alia*, what defines them to be 'rational'. Part of the task of coming up with a maximally coherent and unified set of desires is coming up with a set that would be converged upon by other rational creatures who too are trying to come up with a maximally coherent and unified set of desires; each rational creature is to keep an eye out to her fellows, and to treat as an aberration to be explained, any divergence between the sets of desires they come up with through the process of systematic justification.[5, 6]

---

4 See especially Williams's discussion of the Owen Wingrave example (1980: 106–11).

5 Compare Philip Pettit on rule-following (1993, especially 96–97).

6 The claim is not that on the non-relative conception of reasons the existence of reasons-in-the-actual-world presupposes a convergence in the desires of fully rational creatures in the actual world. For this is itself a relative conception of reasons: reasons are *world*-relative. The non-relative conception really is *non*-relative. It claims that there is a convergence in the desires that all possible creatures would have, so long as those creatures are fully rational, whether those creatures exist in the actual world or not. Angels, ourselves in other possible worlds, the inhabitants of Mars—on the non-relative conception we are all of us supposed to desire the very same thing for the various circumstances we might face, at least insofar as we are rational.

The final question to ask, then, in spelling out our idea of 'full rationality', is whether Williams is right that our ordinary concept of a reason is Humean or anti-Humean. Does our ordinary concept of a reason presuppose skepticism about the scope for reasoned change in our desires? In other words, does it presuppose that there will, or alternatively that there will not, be a convergence in the desires that we would have under conditions of full rationality? If it presupposes that there will not be such a convergence then our concept of a reason is indeed relative, just as Williams says. If it presupposes instead that there will be such a convergence then our concept of a reason is, by contrast, non-relative.

Let me emphasise that we are asking a conceptual question, not a substantive question. We are asking what we mean when we talk of people being fully rational; whether it is part of what we mean by 'rational' that fully rational people converge in their desires, or whether this is no part of what we mean by 'rational'. And note as well that no matter how we answer this question, we do not thereby beg any substantive questions. For example, even if our concept of a reason is itself non-relative—even if our concept optimistically presupposes that we would all converge on the same desires under conditions of full rationality—the world might disappoint us. Entrenched and apparently rationally inexplicable differences in what we desire might make it impossible to believe, substantively, that there are any such non-relative reasons (Smith 1991, 1993, 1994).

Let's, then, confront the conceptual question head on. Is our ordinary concept of a reason relative or non-relative? The relativity of a claim should manifest itself in the way we talk. Consider, for example, the schematic claim 'It is desirable that p in circumstances C'. On the non-relative conception of internal reasons—at least if we abstract away from some complications to be dealt with presently— this claim has a straightforward truth condition: it is desirable that p in C just in case we would all desire that p in C if we were fully rational. There is, then, a sense in which we can talk about rational justification or desirability *simpliciter*. When you and I talk about the reasons that there are for acting, we are therefore talking about the same thing. We are talking about reasons *period*; about the common set of reasons that are appreciable by each of us.

On the relative conception, however, matters are quite different. For in order to give the truth condition of the schematic claim 'It is desirable that p in C' we need first to know from whose perspective the truth of the claim is to be assessed. For while 'It is desirable that p in C' as assessed from A's perspective is true if and only if A would desire that p in C if A were fully rational, 'It is desirable that p in C' as assessed from B's perspective is true if and only if B would desire that p in C if B were fully rational, and so on and so forth. There is thus no such thing as desirability or the considerations that rationally justify *simpliciter*, but only desirability$_A$, desirability$_B$,...; consid-

erations that rationally-justify-from-A's-perspective, rationally-justify-from-B's-perspective,...and so on. If I say to you 'There is a reason for ϕ-ing', and you deny this, we are therefore potentially talking about quite different things: $\text{reasons}_{\text{me}}$ and $\text{reasons}_{\text{you}}$. The question to ask is therefore whether the way in which we talk about reasons for action and the considerations that rationally justify our actions reflects a relative or a non-relative conception of the truth conditions of reason claims.

One reason for thinking that it reflects the non-relative conception comes from the broader context in which the question is being asked. For it is important to remember that we have a whole range of normative concepts: truth, meaning, support, entailment, desirability, and so on. Between them these concepts allow us to ask all sorts of normative questions, questions about what we should and should not believe, say and do. But how many of these other normative concepts are plausibly thought to give rise to claims having relativised truth conditions? As I understand it, none of them do.

Consider our concept of support, by way of example. It seems quite implausible to suppose that the truth of claims about which propositions support which others is implicitly relative to the individual; that when A says 'p supports q' and B says 'p does not support q' they are potentially talking about quite different things: that A is talking about what $\text{supports}_{\text{A}}$ q and B is talking about what $\text{supports}_{\text{B}}$ q, for instance. For if this were the case then we should expect to find that we are sometimes able to dissolve apparent disagreements by finding that both parties are speaking truly. It should be permissable for B to say 'A said "p supports q" and what she said is true, but p does not support q'. However it is a striking feature of our talk about which propositions support which others that we *never* dissolve apparent disagreements in this way. Propositions have normative force *simpliciter*, not just normative-force-relative-to-this-individual or -relative-to-that. When one individual says 'p supports q' and the other says 'p does not support q' they thus express their disagreement about whether p supports q in a *non-relative* sense.

If our concept of desirability were implicitly relativised, then, it seems that this would mark a significant difference between this concept and our other normative concepts. We should expect to find that with claims about what is desirable, unlike claims about which propositions support which, we *are* able to dissolve apparent disagreements in the way just described. But do we find this?

It might be thought that we do. After all, aren't there all sorts of familiar cases in which we say things like 'That may be a reason for you, but it isn't for me', 'Desirable for you maybe, but not desirable for me', and the like? But though there are indeed such cases, it is important to note that the sort of relativity we signal when we say such things is quite different from the kind just described; quite different from the kind of relativity Williams has in mind. For, in the familiar cases, 'That may be a reason for you, but it isn't

for me' signals the fact that there is a relativity built in to the *considerations* that we use to rationally justify our choices. It does not signal the fact that *our concept of a reason* is itself relative to the individual; that there is no such thing as which considerations, relative or not, rationally justify our choices, but only which considerations rationally-justify-relative-to-this-person or rationally-justify-relative-to-that-person. Here, then, we come to the complications abstracted away from earlier.

Sometimes what we have in mind when we say 'That may be a reason for you, but it isn't for me' is that the considerations that rationally justify our choices are, to use Parfit's terms, *agent-relative*, rather than *agent-neutral* (Parfit 1984). Suppose you are standing on a beach. Two people are drowning to your left and one is drowning to your right. You can either swim left and save two, in which case the one on the right will drown, or you can swim right and save one, in which case the two on the left will drown. You decide to swim right and save the one and you justify your choice by saying 'The one on the right is my child, whereas the two on the left are perfect strangers to me'.

In one sense, of course, I may well say 'That may be a reason for you, but it isn't for me'. For if the three people drowning are all perfect strangers to me then, had I been standing on the beach instead of you, I would not have been able to justify the choice of swimming right and saving the one. But in another sense it seems that what is a reason for you may indeed be a reason for me. For if I had been standing on the beach instead of you, and if the one on the right had been my child—that is, if my circumstances had been in all crucial respects *the same* as your's—then surely I too would have been able to justify the choice of swimming right and saving the one by saying 'The one on the right is my child'. Indeed, if we think that a parent who fails to save her child in such circumstances fails to act on a reason available to her—as it seems to me that we do—then we are in fact obliged to say this; obliged to assume the non-relative conception of internal reasons.

What this sort of example shows is therefore that, even if reasons are non-relative in the crucial sense at issue here, among the considerations that may rationally justify our choices are both considerations that are properly given a *de dicto* formulation and considerations that are properly given a *de se* formulation (see also Lewis 1989). That is there are both *de dicto* and *de se* internal reasons. We can each express the content of the *de dicto* reason relevant in this case by using the words 'There is a reason to save people quite generally'. And we can each express the content of the *de se* reason by using the words 'There is a reason to save my child in particular'. In these terms what is a reason for you, in this case, is not a reason for me in the sense that, if it had been me standing on the beach rather than you, and if the same people had been drowning, then the only consideration that would have been relevant to my choice is the *de dicto* reason. The *de se* reason would not have been rel-

evant to my choice because the people who are in fact drowning are all perfect strangers to me. But in another sense what is a reason for you is indeed a reason for me. For if I had been standing on the beach and the one person on the right had been my child, as the one on the right is your child, then both the *de se* and the *de dicto* reason would have been relevant to my choice in just the way they are both relevant to your's.

I said that this sort of relativity is entirely different from the kind that Williams has in mind and it should now be plain why this is so. For, in terms of the analysis, even if some of the considerations that rationally justify our choices are relative because *de se*, the existence of such *de se* reasons may still require a convergence in the desires that we would all have if we were fully rational. That is, the existence of reasons with *de se* contents may still require that, under conditions of full rationality, we would each have desires whose contents we would express by using words like 'to help my children', 'to promote my welfare', and the like. The mere existence of *de se* reasons is thus quite different from the relativity Williams has in mind. For his claim is that reasons are relative in the sense of requiring no such convergence; that the fact that my act helps my child may constitute a $\text{reason}_{\text{me}}$ even though the fact that your act helps your child does not constitute a $\text{reason}_{\text{you}}$.

There is another familiar sort of relativity in our claims about the reasons we have as well, a sort that derives from the fact that what we have reason to do is relative to our circumstances, where our circumstances may include aspects of our own psychology. Suppose, for example, that you and I differ in our preferences for wine over beer. Preferring wine, as you do, you may tell me that there is a reason to go to the local wine bar after work for a drink, for they sell very good wine. But then, preferring beer, as I do, I may quite rightly reply 'That may be a reason for you to go to the wine bar, but it is not a reason for me'.

Now while this might initially look like the claim that our reasons are relative to our desires in something like the sense Williams has in mind, it again isn't really. For the crucial point in this case is that a relevant feature of your circumstances is your preference for wine, whereas a relevant feature of my circumstances is my preference for beer. That this is a relevant feature of our circumstances is manifest from the fact that I can quite happily agree with you that if I were in your circumstances—if I preferred wine to beer—then the fact that the local wine bar sells very good wine would constitute a reason for me to go there as well, just as it constitutes a reason for you.

This sort of relativity is thus completely different from the kind that Williams has in mind as well. For, in terms of the analysis, even if an agent's preferences may enter into a specification of the circumstances that she faces it might still be the case that whether or not she is rationally justified in taking her own preferences into account, and the way in which she is

justified in taking them into account, if she is, depends on whether fully rational agents would all converge on a desire which makes the preferences she in fact has relevant in that way to her choice. In this case, for example, it may be crucial that, under conditions of full rationality we would all converge on a desire to satisfy whatever preferences we might have (perhaps within limits) in deciding where to go for a drink after work.[7] The fact that in rationally justifying our choices our preferences may sometimes be a relevant feature of our circumstances thus does nothing to support Williams's view that our reasons are relative; does nothing to support the view that really there are only the considerations that rationally-justify-relative-to-this-person or rationally-justify-relative-to-that.

In order to find support for the sort of relativity Williams has in mind, we therefore need to look for cases in which it is permissable to make much more radically relativised claims about what there is reason to do. But in fact, as far as I can tell, we find no such claims. Suppose someone tells me that she has a reason to take a holiday and that I think I would have no reason to take a holiday in the circumstances she faces. Provided we have taken proper account of the *de se* considerations that might be relevant to her choice, and provided we have taken proper account of the way in which her preferences may constitute a relevant feature of her circumstances, it seems that I straightforwardly disagree with her about the rational justifiability of her taking a holiday in the circumstances she faces, a disagreement I can express by saying 'She thinks that there is a reason to take a holiday in her circumstances, but there is no such reason'. If she cites a consideration in support of her taking a holiday that I think fails to justify, then I do not conclude that it may justify-relative-to-her, though not justify-relative-to-me, I conclude that it fails to justify *simpliciter*.

The point is important, for it suggests that when we talk about reasons for action we quite generally take ourselves to be talking about a common subject matter: reasons *period*. We are thus potentially in agreement or disagreement with each other about what constitutes a reason and what doesn't. This is why, when we find ourselves in disagreement—as for example in the case of disagreement about whether or not there is a reason to take a holiday in certain circumstances—we always have the option of engaging in argument in the attempt to find out who is right and who is wrong. Other people's opinions about the reasons that there are thus constitute potential challenges to my own opinions. I have something to learn about myself and my own assessment of the reasons that there are by finding out about others and

---

7 Note that the preferences we have are not always a relevant feature of our circumstances. If I just so happen to prefer kicking the cat to leaving it sleep in peace, my fully rational self might want that I do not kick the cat despite my preference. For relevant discussion of this point, and the relevance of actual desires to the desirability or justifiability of our actions generally, see Pettit and Smith 1990, 1993, forthcoming.

their assessment. This is why books and films are so engaging. All of this is flat out inconsistent with the claim that our concept of a reason for action is quite generally relative to the individual; that it typically means $\text{reason}_{\text{me}}$ out of my mouth, $\text{reason}_{\text{you}}$ out of your's, $\text{reason}_{\text{her}}$ out of her's and so on. It suggests rather that our concept of a reason is stubbornly non-relative.

Indeed, it seems to me that we have no choice but to think this. For if reasons were indeed relative then mere reflection on that fact would itself suffice to undermine their normative significance. In order to see why, remember that on the relative conception it turns out that, for example, the $\text{desirability}_{\text{me}}$ of some consideration, p, is entirely dependent on the fact that *my* actual desires are such that, if *I* were to engage in a process of systematically justifying *my* desires, weeding out those that aren't justified and acquiring those that are, a desire that p would be one of the desires *I* would end up having. But what my actual desires are to begin with is, on this relative conception of internal reasons, an entirely *arbitrary* matter, one without any normative significance of its own. I might have had any old set of desires to begin with, even a set that delivered up the desire that not p after a process of systematic justification! The $\text{desirability}_{\text{me}}$ of the fact that p thus turns out to be an entirely arbitrary fact about p. But this is surely a *reductio*, as *arbitrariness* is precisely a feature of a consideration that tends to undermine any normative significance it might initially appear to have. Internal reasons on the relative conception are thus without normative significance (Darwall 1983, 218–39; Smith 1989; Darwall, Gibbard and Railton 1992). And if this is right then it follows that *relative* internal reasons are not *reasons* at all.

On the non-relative conception, by contrast, reflection on our concept of desirability reveals no such arbitrariness. For on that conception everyone is supposed able to reason themselves towards the same desires if they engage in a process of systematic justification of their desires, and they are supposed able to do so precisely because the task of systematic justification is *inter alia* a matter of finding desires that can be shared by their fellow rational creatures. Which desires *I* would end up with, after engaging in such a process, thus in no way depends on what *my* actual desires are to begin with, because reason itself determines the content of our fully rational desires, not the arbitrary fact that we have the actual desires that we have. On the non-relative conception, reflection on the concept of desirability thus leaves the normative significance of facts about what is desirable and undesirable perfectly intact.

This, then, is the final element in our account of what it means when the internalism requirement tells us that the desirability of an agent's ϕ-ing in certain circumstances C depends on whether or not she would desire that she ϕs in C if she were 'fully rational'. Fully rational agents *converge* in their desires about what is to be done in the various circumstances they might face. Of course, the mere fact that a convergence in the hypothetical desires of fully rational creatures is required for the truth of internal reason claims does noth-

ing to guarantee that such a convergence is forthcoming. In defending the non-relative *conception* of internal reasons we have said nothing to suggest that, *substantively*, there are any such reasons. But what we have said does suggest that, in order to discover whether there are any such reasons, and if so what they are, we have no alternative but to give the arguments and see where they lead. Substantive convergence is always assumed available, in so far as we converse and argue about the reasons that we have. But whether or not this assumption is true is always *sub judice*; something to be discovered by the outcome of those very conversations and arguments; something that will emerge when we see where our attempts to systematically justify our desires lead us.

3. The advice model and the appeal of the internalism requirement

So far I have argued that the internalism requirement on reasons is best understood in terms of the 'advice' model, rather than the 'example' model, and I have argued that reasons, understood in terms of the 'advice' model, are best thought of as being non-relative, rather than relative. The two points are related, of course. For I have argued that it is only if we think of reasons on the 'advice' model, and it is only if we think of reasons as being non-relative, that we can properly account for the normative significance of reason claims. However the most important question about the internalism requirement remains yet to be answered. Why exactly should we accept the internalism requirement in the first place? Why shouldn't we think, instead, that reasons have nothing to do, constitutively, with the desires of fully rational agents, as I have defined the idea of 'full rationality'? The answer is that the internalism requirement on reasons enables us to solve an otherwise disturbing puzzle about the role of deliberation in the production of action. Let me begin by explaining the puzzle.

Hume taught us that desires and means-end beliefs each play an essential role in the explanation of action (Smith 1987). Suppose, for example, that all we know about someone is that she believes that if she flicks a particular switch the light will go on and that if she refrains the light will stay off. Then, so far, we have no more reason to suppose that she will flick the switch than refrain. Whether she will flick or refrain must therefore depend on something else about her beyond her beliefs about the way the world is. And indeed it does. It depends on what she happens to desire. Does she desire the causal upshot of flicking the switch, the light's being on, or the causal upshot of refraining from doing so, the light's being off? If the former, then she will flick the switch; if the latter, then she will refrain. Desires are thus essential for the explanation of action. But so are beliefs as well. For if all we know about someone is that she desires the light to be on then, again, so far we have no more reason to suppose that she will flick the switch than that

she will refrain. For whether she will flick the switch or refrain depends on whether she believes the light's being on is the causal upshot of flicking or refraining. To sum up: beliefs alone are unable to motivate action, for beliefs can only motivate action in conjunction with a separate desire; but desires alone are also unable to motivate action, for desires can only motivate action in conjunction with a separate means-end belief.

Compelling though this Humean story of how we explain action is, it presents us with a disturbing puzzle about the role of deliberation in the production of action. For it seems undeniable that we sometimes deliberate in order to find out what we are rationally justified in doing: that is, we sometimes deliberate in order to form beliefs about what it is desirable to do. And it also seems undeniable that we sometimes act upon the outcome of those very deliberations: that is, we sometimes do what we do because we believe that doing so is desirable. But the Humean story about how we explain action seems to leave no room for these undeniable facts. For the belief that it is desirable to act in a certain way is not itself a desire, it is a belief, and so whether or not we happen to act in accordance with this belief, given the Humean story about how we explain actions, must depend entirely on whether we just so happen to have a desire to act in that way, or just so happen to have some other desire which can combine with this belief to yield a desire to act in that way.[8] On Hume's account of the matter it thus appears to be a massive fluke, an inexplicable miracle of nature, that our desires match our beliefs about what it is desirable to do to the extent that they do. For there is nothing in the nature of our evaluative beliefs to explain why this should be the case. What is needed is an extra desire, an extra desire we are not rationally required to have.

Here we see the real appeal of the internalism requirement. For it promises to explain how it can be that our beliefs about what we are rationally justified in doing play a proper causal role in the genesis of our actions, and it promises to do so while leaving Hume's story about the way in which actions are explained largely intact. In order to see why, consider again what the requirement tells us about the content of our evaluative beliefs, at least on the advice model.

When I believe that it would be desirable to ϕ in certain circumstances C, the internalism requirement tells us that my belief has the following content: that I would desire that I ϕ in C if I were fully rational. But now, if indeed I do believe this, and if I believe that I am in circumstances C, then surely the only rational thing for me to desire is to ϕ. For a psychology that includes

---

8 For example, it might be supposed that when we deliberate we *de facto* have a desire to do what we believe it is desirable to do. I will have more to say about this in footnote 10. The point here is simply that the Humean must regard it a happy accident that we all just so happen to have such a desire. For the Humean cannot agree that such a desire is itself required by reason.

both the belief that I would desire that I ϕ in C if I were fully rational—that is, the belief that I would have that desire if my desires formed a maximally coherent and unified set—*and* the desire that I ϕ in C is itself a more coherent and unified psychology than one that includes the belief that I would desire to ϕ in C if I were fully rational and yet *lacks* the desire to ϕ in C. Coherence and unity are thus on the side of a *match* between the content of our evaluative beliefs and our desires.

Here is another way of putting the same point. What would an agent's fully rational self want her less than fully rational self to desire in circumstances in which her less than fully rational self believes that she would desire to ϕ in C if she were fully rational? On the plausible assumption that the agent's fully rational self desires that the psychology of her less than fully rational self is as coherent as possible she will want her less than fully rational self to desire that she ϕs in C. It thus follows that it is desirable for an agent to desire that she ϕs in C in circumstances in which she believes that it is desirable that she ϕs in C. Agents thus quite generally have a reason to desire in accordance with their evaluative beliefs.[9]

But if this is right then it follows that in *rational* creatures at least—that is, in those who do not manifest the form of unreasonableness or irrationality just described, those who are sensitive to the facts about what they have reason to desire—we would therefore expect there to be a causal connection between believing that it is desirable to act in a certain way and desiring to act in that way. That is, given the internalist account of the content of our evaluative beliefs, we would expect a rational deliberator's evaluative beliefs to cause her to have matching desires in much the same way, and for much the same reason, as the rational thinker's beliefs that p and that $p \rightarrow q$ cause her to believe that q. For the psychological states of rational deliberators and thinkers connect with each other in just the way that they rationally should. In this way, then, the internalism requirement can thus underwrite not just the rationality of desiring in accordance with our evaluative beliefs, but also the effectiveness of our evaluative beliefs in bringing about these desires in those who are rational.[10]

---

9 It is, of course, consistent to claim both that: (i) it is desirable that an agent desires to ϕ in C *in circumstances in which she believes that it is desirable to ϕ in C*, and (ii) it is not desirable that an agent desires to ϕ *in circumstances C*. For whereas (i) tells us what an agent's fully rational self would want her less than fully rational self to desire in one set of circumstances, (ii) tells us what her fully rational self would want her less than fully rational self to desire in another, quite different, set of circumstances. The point is important, as it serves to explain why certain theories of reasons for action are properly thought to be *self-effacing* (Smith 1994, chapter 5 footnote 2).

10 Note that the externalist who tries to explain the effectiveness of deliberation by positing an extra desire to do what we believe desirable (see footnote 8) has an explanation that is inferior to the internalist's explanation just given in two respects. First, since the externalist claims that the extra desire to do what we believe desirable is itself rationally optional, he is committed not just to the view that it is a miracle of nature, a

Note that the explanation just given is simply unavailable if we reject the internalism requirement. For on an externalist conception of reasons, the reasons we have are not themselves defined in terms of what we would desire if our psychology exhibited maximal coherence and unity. Without inquiring further into what exactly the content of a reason claim on such a conception is we can therefore already see that there is no reason to expect that a psychology which pairs a belief that there is reason to ϕ in circumstances C with a desire to do something other than ϕ in C will exhibit less in the way of coherence and unity than a psychology that pairs that belief with the desire to ϕ in C. It thus appears that externalists will be unable to explain why it is rational to desire in accordance with our beliefs about the reasons that we have.

Note also that the explanation just given presupposes not just the internalism requirement, but the internalism requirement understood in terms of the advice model. For if we interpret the internalism requirement in terms of the example model, the argument just given simply fails to go through at the crucial point. Suppose, for instance, that you believe your fully rational self would desire to ϕ in the circumstances she faces; that this is the example she would set for you in her own world. Why should this have any effect at all on what you desire to do in the circumstances you face? If your circumstances are quite unlike her's, then you can quite rationally acknowledge her example, and be impressed by it, while still being left entirely unmoved. Coherence and unity do not argue in favour of acquiring a desire like her's because her example—marvelous though it is in the circumstances in which *she* finds *herself*—doesn't engage with the circumstances in which *you* find *yourself*. This is not the case if instead we interpret the requirement in terms of the advice model. For then what you have to believe is that your fully rational self would want your less than fully rational self to ϕ in the circumstances your less than fully rational self actually faces. Your fully rational self's advice engages with your predicament because it is precisely tailored to it. You may

---

massive fluke, that so many of us just so happen to have such a desire, but also to the view that if someone just so happened to lack such a desire, that would not itself suffice to show that that person was irrational. By contrast the internalist has a principled reason for insisting that someone who lacks a desire to ϕ while believing that ϕ-ing is desirable is *as such* irrational. Second, the externalist who posits a quite general desire to do what is desirable must think that if we end up desiring to, say, ϕ in C, as a result of coming to believe that it is desirable to ϕ in C, then the desire to ϕ in C must itself, of necessity, be an *instrumental* desire. The externalist must therefore hold that deliberation never produces a non-instrumental desire to do what we believe desirable, where this is read *de re* rather than *de dicto*. The only thing we desire to do non-instrumentally, when we deliberate, is what it is desirable to do, where this is read *de dicto* rather than *de re*. This seems to me to be an extremely implausible claim. Indeed, as I have argued elsewhere, it seems to constitute a *reductio* of externalism (1994: Chapter 3). The internalist, by contrast, has an explanation of how the belief that it is desirable to ϕ in C generates a desire to ϕ in C that is perfectly consistent with the claim that the resulting desire to ϕ in C is *non-instrumental* in character.

still say 'So what?', of course, but if you do you simply reveal that you are unable to accept good advice; you reveal the extent to which your psychology fails in terms of norms of coherence and unity that define a systematically justified psychology. You thus simply betray your own irrationality.

Here, then, we see the real appeal of the internalism requirement. It offers us an explanation of how and why our evaluative beliefs come to play a proper causal role in the production of our desires, an explanation that leaves the Humean's claim that intentional actions are themselves the product of desires and means-end beliefs perfectly intact. The crucial idea, to repeat, is that given the content of an agent's evaluative beliefs—that is, given the internalism requirement—the desires that the Humean rightly supposes play a causal role in the genesis of intentional actions will themselves be caused by the agent's evaluative beliefs to the extent that she is a rational deliberator. The Humean's account has thus been supplemented, not replaced.

Conclusion

My aim in this paper has been to answer three questions. How exactly is the internalism requirement on reasons to be understood? What does it tell us about the nature of reasons? And where-in lies its appeal?

As regards the first question, I have argued that the content of the internalism requirement is best captured by what I have called the 'advice' model rather than the 'example' model. According to the advice model, the desirability of an agent's ϕ-ing in certain circumstances C is fixed by whether or not her fully rational self would advise her less than fully rational self to ϕ in the circumstances that she, the less than fully rational self, faces: that is, in circumstances C. The idea is not that the desirability of an agent's ϕ-ing in C is fixed by the example her fully rational self would set for her less than fully rational self by her own behaviour in her own world. Thus, even though the requirement is concerned with the *desires* of a fully rational agent, it is crucially not concerned with the *motivations* of a fully rational agent.

As regards the second question, I have argued that the substantive content of the internalism requirement depends on the way in which we understand the key idea of having certain desires under conditions of 'full rationality'. My claim has been that it is part of our concept of 'full rationality' that fully rational agents are those who have a systematically justifiable set of desires, where this idea is to be cashed out in terms of having a psychology that is maximally coherent and unified, and where it is presupposed that the maximally coherent and unified set of desires any one particular fully rational agent would come up with is exactly the same as the maximally coherent and unified set of desires any other rational agent would come up with. The internalism requirement is thus best understood as offering us a non-relativistic, rather than a relativistic, conception of reasons.

Finally, as regards the third question, I have argued that, given our answers to the earlier two questions, the appeal of the internalism requirement is easy to understand. For it allows us to see that though the Humean is right that all *actions* are caused by desires, in rational deliberators at least, the *desires* that cause an agent's actions may themselves be caused by her evaluative beliefs. The internalism requirement thus enables us to assign a proper causal role to an agent's beliefs about the rational justifiability of her actions when she deliberates.

For all I have said it of course remains an open possibility that there are no internal reasons—and hence that there are no reasons for action at all. After all, the mere fact that our concept of a reason presupposes that fully rational creatures would converge in their desires does nothing to show that such a convergence is forthcoming. But that is no objection to what has been said here. For my aim has not been to argue that there are any reasons, it has rather been to articulate the conceptual framework in which debates about what our reasons are, if there are any, can sensibly take place.[11]

REFERENCES

Brandt, Richard 1979: *A Theory of the Good and the Right*. Oxford University Press.

Daniels, Norman 1979: 'Wide Reflective Equilibrium and Theory Acceptance in Ethics' *Journal of Philosophy*. 256–82.

Darwall, Stephen 1983: *Impartial Reason*. Cornell University Press.

_____ Allan Gibbard and Peter Railton 1992: 'Toward *Fin de siecle* Ethics: Some Trends', *Philosophical Review*. 115–89.

Korsgaard, Christine 1986: 'Skepticism about Practical Reason', *Journal of Philosophy*. 5–25.

Lewis, David 1989: 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society* Supplementary Volume. 113–37.

Johnston, Mark 1989: 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society* Supplementary Volume. 139–74.

Parfit, Derek 1984: *Reasons and Persons*. Oxford University Press.

Pettit, Philip 1993: *The Common Mind*. Oxford University Press.

_____ and Michael Smith 1990: 'Backgrounding Desire', *The Philosophical Review*. 565–92.

---

[11] An earlier version of this paper was presented at 'Internal and External Reasons', a symposium held at the Pacific Division APA meetings in Los Angeles, April 1994. I would like to thank Stephen Darwall for the many useful suggestions and observations he made as commentator on that occasion, suggestions and observations that have helped me greatly improve the paper. I also received useful advice from John Broome, David Copp, Frank Jackson, Douglas Maclean, Kevin Mulligan, Philip Pettit, Denis Robinson, Holly Smith, Galen Strawson, Anita Superson, Sigrun Svavarsdottir, David Velleman and Susan Wolf. The second section of the paper draws on material that appears in Chapter Five of *The Moral Problem* (Basil Blackwell, 1994).

_____ and Michael Smith 1993: 'Brandt on Self-Control' in Brad Hooker, ed., *Rationality, Rules and Utility*. Westview Press.
_____ and Michael Smith forthcoming: 'Parfit's P' in Jonathan Dancy, ed., *Parfit and his Critics 2: Reasons*. Blackwell.
Railton, Peter 1986: 'Moral Realism', *The Philosophical Review*. 163–207.
Rawls, John 1951: 'Outline of a Decision Procedure for Ethics', *Philosophical Review*. 177–97.
_____ 1971: *A Theory of Justice*. Harvard University Press.
Shope, Robert K. 1978: 'Rawls, Brandt, and the Definition of Rational Desires', *Canadian Journal of Philosophy*. 329–40.
Smith, Michael 1987: 'The Humean Theory of Motivation', *Mind*. 36–61.
_____ 1989: 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society* Supplementary Volume. 89–111.
_____ 1991: 'Realism' in Peter Singer, ed., *A Companion to Ethics*. Basil Blackwell. 399–410.
_____ 1992: 'Valuing: Desiring or Believing?' in David Charles and Kathleen Lennon, eds., *Reduction, Explanation, Realism*. Oxford University Press. 323–60.
_____ 1993: 'Objectivity and Moral Realism: On the Phenomenology of Moral Experience' in John Haldane and Crispin Wright, eds, *Reality, Representation and Projection*. Oxford University Press. 235–36.
_____ 1994: *The Moral Problem*. Basil Blackwell.
Watson, Gary 1975: 'Free Agency' reprinted in Gary Watson, ed., *Free Will*. Oxford University Press. 1982. 96–110.
Bernard Williams 1980: 'Internal and External Reasons' reprinted in his *Moral Luck*. Cambridge University Press. 1981.