AN OFFPRINT FROM

REDUCTION, EXPLANATION, AND REALISM

Edited by

DAVID CHARLES

AND

KATHLEEN LENNON

CLARENDON PRESS · OXFORD

1992

12

Valuing: Desiring or Believing?*

Michael Smith

1. THE INTENTIONAL AND THE DELIBERATIVE

In general, we can explain intentional action from two quite different perspectives: the *intentional* and the *deliberative*.[1] Though at first sight the difference between these two perspectives is both intuitive and clear, on reflection the fact that intentional action can be explained from these two perspectives provides us with a puzzle. Let me begin by explaining the two perspectives.

From the intentional perspective, we explain an intentional action by fitting it into a pattern of teleological, and perhaps causal, explanation: in other words, we explain by citing the complex of psychological states that produced the action. Consider my typing these words. We explain this action from the intentional perspective when we cite my desire to produce a paper and my belief that I can do so by typing these words, for this is the desire–belief pair that teleologically, and perhaps causally, explains my typing these words.

From the deliberative perspective, by contrast, we explain an intentional action in terms of the pattern of deliberation that either did, or could have, produced it. Consider again my typing these words. In deciding whether or not to type these words I reflected on certain facts: that it would be desirable to

* An earlier version of this paper was read at Oriel College, Oxford University, and at the Research School of Social Sciences, Australian National University. In addition to Philip Pettit, I would like to thank Will Barrett, Geoff Brennan, David Charles, David Gauthier, Brad Hooker, Lloyd Humberstone, Julie Jack, Jeanette Kennett, Peter Menzies, Paul Snowdon, Michael Tooley, and J. David Velleman for their many helpful comments

[1] This terminology is introduced in P. Pettit and M. Smith, 'Backgrounding Desire', *Philosophical Review*, 99 (1990). The present paper attempts to answer some of the questions on which we remained deliberately neutral in that paper. I am grateful to Philip Pettit for helpful comments and conversations.

write a paper and that I can do so by typing these words. These are amongst the considerations I actually took into account in deciding what to do before I did it; they give my reasons. Of course, it would be wrong to suppose that we consciously go through such a process of reasoning each time we act. However, even when we don't, we can often reconstruct the pattern of reasoning that could have been explicit in our deciding to do what we did. In this case, *ex post facto* justification takes the form of constructing an 'as if' story, a story that may be more or less close to the truth.

At first sight, then, the distinction between the intentional and the deliberative may seem to involve a difference in *category*. For whereas, from the intentional perspective, we are interested in which *psychological states* of the agent *explain* his actions, from the deliberative perspective we seem to be interested in which *propositions*, from the agent's point of view, *justify* his actions. However, on closer examination, we see that both perspectives commit us to claims about explanation and that they are therefore potentially in conflict.

In order to see that the deliberative perspective is indeed a perspective on explanation, it suffices to note that to imagine otherwise is tantamount to supposing that the connection between what we decide to do, on the basis of deliberation, and what we do do, is altogether contingent and fortuitous. And that is patently absurd. But if deliberation is practical not just in its *content*, but also in its *issue*, then we must suppose that our attitudes towards the propositions that figure in our deliberations also figure in the explanation of what we do. And now the potential for conflict arises.

Consider again our example. If I accept that it is desirable that I write a paper and that I can write a paper by typing these words, then it seems uncontroversial to say that I *value* writing a paper and *believe* that I can write a paper by typing these words. We might say, then, that the attitudes in question are valuing and believing. But now we must ask how this deliberative explanation in terms of *valuing* and believing relates to the intentional explanation of the very same action in terms of *desiring* and believing.

The problematic attitude is valuing. What is valuing? And how does valuing relate to desiring?

2. DEVIANT CASES

An answer to the question 'What is valuing?' must remain faithful to ordinary thought. In this section I describe part of our ordinary thought about the relations between valuing and desiring. I discuss the proper interpretation of ordinary thought in the next section.

Consider the following passage from Ayer's paper 'Freedom and Necessity':

> The kleptomaniac does not go through any process of deciding whether or not to steal. Or rather, if he does go through such a process, it is irrelevant to his behaviour. Whatever he resolved to do, he would steal all the same. And it is this that distinguishes him from the ordinary thief.[2]

Ayer rightly takes it as given that the kleptomaniac steals intentionally; that is, that what he does is explicable in terms of what he wants to do. The problem he highlights is that, though the kleptomaniac's action is therefore explicable from the intentional perspective, it need not be explicable from the deliberative. For there is, he notes, the potential for a gap between what the kleptomaniac 'resolves' to do as a result of deliberating and what he wants to do. This is important, for it suggests a distinction between valuing and desiring. In particular, it suggests that an agent may desire to act in a certain way without valuing acting in that way.

Harry Frankfurt makes a similar point. He has us imagine a heroin addict who:

> hates his addiction and always struggles desperately, although to no avail, against its thrust. He tries everything that he thinks might enable him to overcome his desires for the drug. But these desires are too powerful for him to withstand, and invariably, in the end, they conquer him. He is an unwilling addict, helplessly violated by his own desires.[3]

The heroin addict certainly wants to take the drug. However, as Frankfurt notes, we can imagine him saying that he 'does

[2] A. J. Ayer, 'Freedom and Necessity', in G. Watson (ed.), *Free Will* (Oxford: Oxford University Press, 1982), 20.

[3] H. Frankfurt, 'Freedom of the Will and the Concept of a Person', in Watson (ed.), *Free Will*, 87.

not "really" want to' take the drug; or even that he 'would rather die than' take it.[4] Here, as elsewhere, talk of what we 'really want' is a surrogate for talk about what we value. Frankfurt's point, like Ayer's, is thus that we may desire to act in a certain way without valuing acting in that way.

Gary Watson makes a related point:

> Consider the case of a woman who has a sudden urge to drown her bawling child in the bath; or the case of a squash player who, while suffering an ignominious defeat, desires to smash his opponent in the face with the racquet. It is just false that the mother values her child's being drowned or that the player values the injury and suffering of his opponent. But they desire these things none the less. They desire them in spite of themselves. It is not that they assign to these actions an initial value which is then outweighed by other considerations. These activities are not even represented by a positive entry, however small, on the initial 'desirability matrix'.[5]

Watson's woman may want to drown her bawling baby, and even do so intentionally. But if she does, she does something that she does not value at all.

The Ayer–Frankfurt–Watson cases all remind us that we may desire something without valuing it. These cases are deviant, to be sure: cases of psychological compulsion, physical addiction, and emotional disturbance. But ordinary thought tells us that these cases exist and are therefore not to be ignored. If we deliberate on the basis of our values, then, to the extent that we may act on the basis of such desires, our actions may fail to reflect our deliberations.

Consider a split of a different kind between what we value and what we desire. In 'Desiring the Bad: An Essay in Moral Psychology' Michael Stocker observes:

> Through spiritual or physical tiredness, through accidie, through weakness of body, through illness, through general apathy, through despair, through inability to concentrate, through a feeling of uselessness or futility, and so on, one may feel less and less motivated to seek what is good. One's lessened desire need not signal, much less be the product of, the fact that, or one's belief that, there is less good to be obtained or produced, as in the case of a universal *Weltshmertz*.

[4] Ibid. 83.
[5] G. Watson, 'Free Agency', in Watson (ed.), *Free Will*, 101.

Indeed, a frequent added defect of being in such 'depressions' is that one sees all the good to be won or saved and one lacks the will, interest, desire or strength.[6]

Where the Ayer–Frankfurt–Watson cases remind us that we may desire something without valuing it, Stocker reminds us that we may value something without desiring it. Again, these cases are deviant: cases of severe 'depression'. But, again, ordinary thought tells us that these cases exist and are not to be ignored. If we deliberate on the basis of our values, then, to the extent that we may act on the basis of desires which do not adequately reflect our values, our actions may fail adequately to reflect our deliberations.

3. VALUES AND NORMATIVE REASONS

It is one thing to describe cases in which ordinary thought tells us there is a gap between deliberation and action, quite another to have a philosophically plausible interpretation of that gap. I attempt to provide such an interpretation in this section.

Note that we ordinarily distinguish two senses in which we can be said to have a reason for action.[7] The first is the sense in which we are happy to acknowledge that to do something intentionally is simply to do that thing for a reason. Here our talk of reasons is talk about the psychological states that motivate what we do, the complex of psychological states that teleologically, and perhaps causally, explain our actions. Or rather, and more accurately since we need not act on our reasons, such talk of reasons is talk about psychological states with the *potential* to motivate or explain behaviour. Let's call these our 'motivating' reasons.

In the second sense, however, we say we have a reason to do all and only those things for which we can construct a certain sort of *justification*. Justifications may, of course, be of quite different kinds. An action may be judged according to

[6] M. Stocker, 'Desiring the Bad: An Essay in Moral Psychology', *Journal of Philosophy*, 76 (1979), 744.

[7] M. Smith, 'The Humean Theory of Motivation', *Mind*, 96 (1987), §2. Note that what follows constitutes a correction of some of what I say there.

standards of rationality, morality, the law, etiquette, and perhaps according to other standards as well. If we are not to beg any questions, we should therefore be prepared to admit that each of these may give rise to reason claims, though such claims may not, of course, be autonomous—one kind of reason may reduce to another.[8] Let's call all these our 'normative' reasons.

When we explain an action from the intentional perspective, we are clearly concerned with our motivating reasons, for we are concerned with the psychological states that teleologically, and perhaps causally, explain our actions. And when we explain an action from the deliberative perspective, we are clearly concerned with our normative reasons, for we are concerned with our justifications for acting in the different ways in which we might act. But what *kinds* of justification concern us in so far as we occupy the deliberative perspective?

Certainly, we may appeal to quite different kinds of justificatory consideration when we deliberate. We may measure the alternatives against standards of morality, etiquette, the law, and perhaps against other standards as well. But, at bottom, when we have to decide how to act, we must weigh these different considerations against each other. And when we do that we seem to be trying to decide what the most reasonable or rational thing to do is. Deliberation thus seems to give a privileged place to our normative reasons where justification is judged according to standards of rationality. For the fact that some alternative is required by morality, the law, or etiquette matters in deliberation only to the extent that it is reasonable or rational to act on these requirements.

Indeed, we might now say that normative reasons, where justification is judged according to standards of rationality (and from here on I will omit this qualification), *are* the values to which we appeal when we deliberate. When we talk about

[8] I discuss the relationship between what is rationally required and what is morally required in 'Reason and Desire', *Proceedings of the Aristotelian Society*, 88 (1988); 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society*, Supp. Vol. 63 (1989); 'Realism', in P. Singer (ed.), *Companion to Ethics* (Oxford: Blackwell, 1991); 'Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience', in C. Wright and J. Haldane (eds.), *Reality, Representation and Projection* (Oxford: Oxford University Press, forthcoming).

deliberating on the basis of our values, all we mean is that we take account of the different ways in which it would be reasonable or rational to act. And here we must emphasize the 'different', for one course of action may, of course, be *more* or *less* reasonable than another. Our concern from the deliberative perspective is with *pro tanto* reasons; with values that may be weighed one against the other; with finding out what is the most reasonable or rational way to act.

If this is right, and if, as seems plausible, motivating reasons are constituted by desires, then we have the following rough equivalences:

(1) *A* has a normative reason to ϕ iff *A*'s ϕing is valuable;
(2) *A* accepts that he has a normative reason to ϕ iff *A* values ϕing;

and

(3) *A* has a motivating reason to ϕ iff *A* desires to ϕ.

And, given these rough equivalences, we are now in a position to offer a philosophically plausible interpretation of the deviant cases served up by ordinary thought.

What is the significance of these deviant cases? The answer seems evident. They simply chart the different ways in which normative reasons and motivating reasons may *come apart*. Thus the Ayer–Frankfurt–Watson cases remind us that we may have a motivating reason to do something we have no normative reason to do; something for which we cannot, and even accept that we cannot, provide a rational justification. And the Stocker cases remind us that we may accept that we have a normative reason to do something, and perhaps even have such a reason, without having a corresponding motivating reason: that is, that we may accept that we have a rational justification for acting in a certain way and yet be entirely unmoved by that fact.

Psychological compulsions, physical addictions, and various emotional disturbances can cause us to desire to do what we cannot rationally justify and fail to desire to do what we can rationally justify. So says ordinary thought. In giving an account of the relationship between valuing and desiring we must respect ordinary thought on this score.

4. THE PUZZLE

I said at the outset that the fact that intentional action can be explained from both the intentional and the deliberative perspectives provides us with a puzzle. We are now in a position to explain that puzzle more fully.

To the extent that reflection on our normative reasons moves us to act—that is, to the extent that we are effective deliberators—accepting that we have certain normative reasons must be bound up with having corresponding motivating reasons. But the deviant cases remind us that our desires may come apart from the normative reason claims we accept—that is, that we may be ineffective deliberators. The puzzle, then, is to explain how it can be that accepting normative reasons can both be *bound up* with having desires and yet *come apart* from having desires. In other words, the problem is to explain how deliberation on the basis of our values can be practical in its issue *to just the extent that it is*.

The puzzle here is a deep one, traceable to the view of human psychology we have inherited from Hume.[9] For, according to Hume, there are two main kinds of psychological state, beliefs and desires, utterly distinct and different from each other. Unfortunately, however, Hume's account of belief and desire seems to leave no room for the idea that deliberation on the basis of our values is practical in its issue to just the extent that it is. Let me explain why.

On the one hand, Hume held that there are *beliefs*, states that purport to represent the way the world is. Since our beliefs purport to represent the world, they are subject to rational criticism: specifically, they are assessable in terms of truth and falsehood according to whether or not they succeed in representing the world to be the way it really is.

On the other hand, however, there are also *desires*, states that represent how the world is to be. Desires are unlike beliefs in that they do not even purport to represent the way the world is. They are, in Hume's terms, 'original existences'.[10] Desires are therefore not assessable in terms of truth and

[9] See esp. David Hume, *A Treatise of Human Nature* (Oxford: Clarendon Press, 1978), esp. II. iii. 3.

[10] *A Treatise of Human Nature*, 415.

falsehood. Indeed, according to the Hume, and contrary to what we have just seen, our desires are, at bottom, not subject to any sort of rational criticism at all.[11] (We will return to this point later.)

The Humean view of human psychology is important because it provides us with a model for understanding human action: indeed, it underwrites the intentional perspective. Human action is, according to this view, produced by a combination of the two kinds of psychological state. Crudely, our beliefs tell us how the world is, and thus how it has to be changed, so as to make it the way our desires tell us it is to be. A desire without a belief would be ignorant of what change has to be made in the world so as to realize its content. A belief without a desire is simply an inert representation of how things are. An action is thus the product of these two forces: a desire representing the way the world is to be and a belief telling us how the world has to be changed so as to make it that way.

The puzzle the deliberative perspective presents should now be clear. We are to suppose that deliberating is practical in its issue to the extent that we act in the ways we do because we value what we value, and that deliberating fails to be practical in its issue to the extent that our values are irrelevant to the determination of what we do. But what is it to *value* something? That is, equivalently, what is it to accept a normative reason to do something? Is it a matter of *believing* or *desiring*? We face a dilemma.

If valuing is a matter of believing, then, given the Humean view of human psychology, it is difficult to see how anything we do intentionally could be done because we value what we value. For our beliefs cannot produce actions; they are simply inert representations of how things are. And if valuing is a matter of desiring then though it is now clear how we can act because of our values, given the Humean view, it is difficult to see how there could be the requisite *gap* between what we value and what we desire.

What is at issue is thus the very coherence of the idea that

[11] I discuss Hume's account of the ways in which we may rationally criticize desires in 'Reason and Desire'.

deliberation on the basis of our values is practical in its issue to just the extent that it is. In the remainder of this paper I therefore consider the two alternatives: that valuing is desiring and that valuing is believing.

5. DAVIDSON ON VALUING AS DESIRING

A reduction of valuing to desiring is proposed by Donald Davidson:

> The natural expression of . . . [an agent's] desire [say, the desire to improve the taste of the stew] . . . is . . . evaluative in form; for example, 'It is desirable to improve the taste of the stew'. We may suppose that different pro attitudes are expressed with other evaluative words in place of 'desirable'.
>
> There is no short proof that evaluative sentences express desires and other pro attitudes in the way that the sentence 'Snow is white' expresses the belief that snow is white. But the following considerations will perhaps help show what is involved. If someone who knows English says honestly 'Snow is white', then he believes snow is white. If my thesis is correct, someone who says honestly 'It is desirable that I stop smoking' has some pro attitude towards his stopping smoking. He feels some inclination to do it; in fact he will do it if nothing stands in the way, he knows how, and he has no contrary values or desires. Given this assumption, it is reasonable to generalize: if explicit value judgements represent pro attitudes, all pro attitudes may be expressed by value judgements that are at least implicit.[12]

Davidson succeeds in making it clear how we get ourselves to act by deliberating. But the cost of his doing so is a distortion in the extent to which deliberation on the basis of our values is practical in its issue. For, according to Davidson, it is impossible for an agent to act on a desire without acting on a normative reason claim he accepts. This is impossible because a desire is appropriately expressed *in* a normative reason claim. The question is therefore whether we are obliged to accept his reduction.

In so far as Davidson offers us an argument for his reduction

[12] D. Davidson, 'Intending', in D. Davidson, *Essays on Actions and Events* (Oxford: Clarendon Press, 1980), 86.

of valuing to desiring, it seems to me easy to see where he goes wrong. In short, he incorrectly assumes that a feature of *rational* evaluators is a feature of *all* evaluators. To be sure, a rational evaluator who says honestly 'It is desirable that I stop smoking' has some pro attitude towards stopping. But it does not follow that someone may not honestly say 'It is desirable that I stop smoking' and have no inclination to stop, and nor does it follow that someone who has some inclination to smoke may none the less be unable honestly to say 'It is desirable that I smoke'. All that follows is that, if either of these is possible, the agents in question are not *rational* evaluators. Davidson needs to rule out these forms of irrationality if his reduction is to secure conviction. And that seems no ordinary task.

In fact, however, in his earlier work on the relationship between reasons and actions, Davidson does offer us an explicit argument for the claim that motivating reasons bring with them the acceptance of corresponding normative reason claims. In 'Actions, Reasons, and Causes', for example, he tells us:

> In the light of a primary reason . . . the agent is shown in his role of Rational Animal. Corresponding to the belief and attitude of a primary reason for an action [i.e. its cause] . . . we can always construct (with a little ingenuity) the premises of a syllogism from which it follows that the action has some (as Anscombe calls it) 'desirability characteristic'. Thus there is a certain irreducible—though somewhat anaemic—sense in which every rationalization justifies: from the agent's point of view there was, when he acted, something to be said for the action.[13]

The idea that a rationalization reveals the agent in his role of rational animal is a central theme in Davidson's work. For elsewhere, expanding on the theme of 'Actions, Reasons, and Causes', he notes that a rationalization must do more than merely provide us with a causal explanation of an action in terms of a desire–belief pair, for the 'problem of wayward causal chains' shows us that it must causally explain the action

[13] D. Davidson, 'Actions, Reasons, and Causes', in Davidson, *Essays on Actions and Events*, 9. The sense is 'anaemic' because the agent need not think that his action is desirable *all things considered*; the justification in question may be merely *pro tanto*. And this is precisely the claim to which I object. For ordinary thought tells us that even this 'anaemic' claim is too strong.

in the right way.[14] And, seeing no way in which this idea can be analysed in terms that do not take as primitive the idea that rationalizations render actions rationally intelligible, Davidson concludes that rationalizations must provide us with justifications. Why? Because, for Davidson, rational intelligibility amounts to seeing an action as having been done for a reason. And seeing an action as having been done for a reason amounts to seeing a rational justification for doing it.

The idea that rational intelligibility entails rational justifiability in some such way is in fact a common one. Thus, for example, Michael Woods has suggested: 'the concept of a reason for an action stands at the point of intersection, so to speak, between the theory of the explanation of actions and the theory of their justification'.[15] Leave out the explanatory dimension and you don't have a reason; for our appreciation of our reasons for action makes a difference to what we do. But leave out the justificatory dimension and you don't have a reason either; that's the lesson of the problem of wayward causal chains, for it teaches us that the idea of an act's having been done for a reason is primitive. Our reasons aren't mere causes, they make our actions rationally intelligible; that is, they show what we do to be rationally justifiable.

Davidson's reduction of valuing to desiring is thus a crucial element in his account of rationalization. For his reduction is supposed to show how it is possible for our desires both to explain and to justify what we do: desires, the psychological states that *causally explain* our actions, are to be thought of as appropriately expressed in evaluations, claims that permit us to *justify* our actions. How are we to reply?

Davidson is, I think, right to insist that rationalization

[14] D. Davidson, 'Freedom to Act', in Davidson, *Essays on Actions and Events*, 77–81. Here is Davidson's example of a 'wayward causal chain': 'A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do it intentionally. It will not help, I think, to add that the belief and the want must combine to cause him to want to loosen his hold, for there will remain the *two* questions *how* the belief and the want caused the second want, and *how* wanting to loosen his hold caused him to loosen his hold' (p. 79).

[15] M. Woods, 'Reasons for Action and Desire', *Proceedings of the Aristotelian Society*, Supp. Vol. 46 (1972), 189.

reveals the agent in his role of rational animal; right that a rationalization must do more than merely provide us with a causal explanation of an action in terms of a desire–belief pair; right that, for it to count as a rationalization, the desire–belief pair cited must causally explain the action in the right way; right that a desire–belief pair causally explains an action in the right way only if it renders the action rationally intelligible, and thus as having been done for a reason, where this idea is not further reducible. The only question is whether he is right that, if a desire–belief pair can do all of this, it must be possible to show that the action in question is rationally justifiable. And the answer is that he is not right.

If irrational action of the kind we have been discussing here is at all possible, then it follows that rational intelligibility can be had *without* rational justifiability. And such actions certainly seem to be possible. Watson's woman who drowns her bawling baby in the bathwater, for example, certainly does something that is rationally intelligible. We do not have here an instance of a wayward causal chain; she acted for a reason. But, as we have seen, what she does is not rationally justifiable, not even by her own lights.

Nor is it difficult to see how such intelligibility can be had in the absence of a justification. For wherever we have an explanation of an intentional action in terms of a desire–belief pair we have an explanation that renders the action intelligible as one that serves a goal had by the agent. Such a teleological explanation, and thus such rational intelligibility, is not available when the action is not intentional; that is, in cases in which a desire–belief pair causally explains an action but not in the right way. The availability of a teleological explanation thus suffices for rational intelligibility, for it suffices to provide us with the motivating reason for which the act was done. However, it does not suffice for rational justifiability. For that we require a normative reason. The woman who drowns her bawling baby in the bathwater acts in a way that serves her goals, and thus her action is teleologically explicable. But what she does is not rationally justifiable, for the goal that she serves is itself unreasonable even by her own lights. She acts on a motivating reason, but she cannot provide herself with a normative reason for acting in that way.

In attempting to reduce valuing to desiring, then, Davidson

ignores this distinction between two ways in which an action can be rendered rationally intelligible. And the fact that he ignores this distinction infects his whole account of rationalization. It makes his claim that rationalizations reveal the agent in his role of 'rational animal' too strong to be credible. We therefore have every reason to reject Davidson's reduction of valuing to desiring in favour of an account that preserves the distinction between teleological explicability and rational justifiability.

6. GAUTHIER ON VALUING AS A MODE OF DESIRING

David Gauthier suggests that valuing is not desiring *simpliciter*, but rather a *mode* of desiring:

> Practical rationality in the most general sense is identified with maximization . . . An objector might agree to identify practical rationality with maximization, but insist that a measure of individual preference is not the appropriate quantity to maximize. It is rational to maximize *value*; the theory of rational choice implicitly identifies value with [a precise measure of preference: i.e. with] . . . utility, but the objector challenges this identification . . . [He might] agree that value is a measure, but insist that it does not measure brute preferences, which may be misinformed, inexperienced, or ill-considered. We shall accept this view in so far as it concerns the manner in which preferences are held.[16]

Though on the surface Gauthier's view is preferable to Davidson's, on examination it is plain that his view too is in conflict with ordinary thought.

According to Gauthier, an agent values a certain outcome only if he desires that outcome *in a certain way*: that is, only if his desires pass certain tests of reflection and experience.[17] Suppose I prefer white wine to red, but without ever having had either a sip or a sniff of red wine. This preference may not pass the test of experience, for, tasting red wine, I might find that I prefer red wine to white. Or suppose I have to decide

[16] D. Gauthier, *Morals by Agreement* (Oxford: Clarendon Press, 1986), 22–3.
[17] Gauthier, *Morals by Agreement*, 29–32.

whether to have white wine or red, and that I choose white, so revealing my preference, but without having given the matter any thought whatsoever. This preference may not pass the test of reflection. For, on reflection, I might have found that I prefer red wine to white.

Though these constraints on the mode of desiring appropriate for valuing are not, as they stand, sufficient to rule out the deviant cases we have described, Gauthier could quite evidently enrich his conception of the appropriate mode. Thus, he might add, valuing is desiring where the desire in question requires no support from a psychological compulsion, a physical addiction, or a state of emotional distress. In this way he could agree with ordinary thought that we may desire an outcome without valuing it.

However, while these amendments would suffice to do that, they would fail altogether to show how we may value an outcome without desiring it. Indeed, if valuing is a mode of desiring at all, as Gauthier suggests, then valuing without desiring becomes a conceptual impossibility. But ordinary thought tells us that valuing without desiring isn't just possible, it is actual. The 'depressions' Michael Stocker describes sap our desires while leaving our evaluative outlooks intact.

The problem with any view according to which valuing is a mode of desiring, then, is that it will only account for the ways in which a desire we have may fail to be rational. It will ignore altogether the ways in which we may fail to be rational in virtue of lacking certain desires. Given that we may accept rational justifications for acting in certain ways and yet, due to depressions, remain unmoved, such cases certainly seem to exist. It therefore follows that we should reject the view that valuing is a mode of desiring.

7. LEWIS ON VALUING AS DESIRING TO DESIRE

It might be thought that a reduction of valuing to desiring is still on the cards. We need simply to reconceive the desire in question in a more radical way. Consider, therefore, David Lewis's reduction of valuing to desiring to desire:

So we turn to desires. But we'd better not say that valuing something is just the same as desiring it. That may do for some of us: those who manage, by strength of will or by good luck, to desire exactly as they desire to desire. But not all of us are so fortunate. The thoughtful addict may desire his euphoric daze, but not value it. Even apart from all the costs and risks, he may hate himself for desiring something he values not at all. It is a desire he wants very much to be rid of. He desires his high, but he does not desire to desire it. He does not desire an unaltered, mundane state of consciousness, but he does desire to desire it. We conclude that he does not value what he desires, but rather he values what he desires to desire.[18]

Lewis's reduction of valuing to desiring to desire avoids the problem facing Davidson's. For an agent may certainly first-order desire other than he second-order desires. And his reduction avoids the problem facing Gauthier's as well. For an agent may certainly second-order desire other than he first-order desires.

In this way, then, it might therefore be thought that Lewis's reduction promises us just the distinction we want between teleological explicability and rational justifiability. An action is teleologically explicable if it is first-order explainable. But in order to be rationally justifiable, an action must be second-order explainable. Let's consider this view in more detail.[19]

The point that emerged in our discussion of Davidson's and Gauthier's reductions of valuing to desiring was that though a *rational* agent desires in accordance with the normative reason claims he accepts, an *irrational* agent may desire other-

[18] D. Lewis, 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society*, Supp. Vol. 63 (1989), 115. In 'Freedom of the Will and the Concept of a Person', Harry Frankfurt also defends the idea that valuing is desiring to desire.

[19] Let's be clear about the question we are asking. Since we have seen no alternative but to identify valuing with accepting a normative reason claim, the question is whether Lewis's reduction of valuing to desiring to desire makes that identification seem plausible. It must be said at the outset, however, that Lewis says nothing to suggest that he would accept such an identification himself. Thus we must not suppose that the discussion that follows constitutes a criticism of Lewis. Rather, his reduction provides us with a useful focus for discussing a view that naturally suggests itself given the failure of Davidson's and Gauthier's reductions. Our criticism of Lewis, at this point, is rather implicit in the argument given earlier for identifying valuing something with taking it to be a reasonable or rational thing to do. Mark Johnston develops this criticism of Lewis at some length in his 'Dispositional Theories of Value', *Proceedings of the Aristotelian Society*, Supp. Vol. 63 (1989), 149–61.

wise. This suggests that normative reasons are subject to the following constraint:

(C1) If an agent accepts that he has a normative reason to ϕ, he rationally should desire to ϕ.

If accepting a normative reason claim is the same as valuing, and valuing is desiring to desire, then it follows that an agent who desires to desire to ϕ rationally should desire to ϕ. But is this right?

It might be thought that it is. Consider once again the Humean picture of the difference between beliefs and desires. The role of a belief is, you will recall, to represent the world. Thus a belief that arises independently of the way the world is is defective, a belief we can rationally criticize. But now consider desires by analogy.

The role of a desire is to make the world be the way it says it is to be. Thus a desire that persists without changing the world in the requisite way is, by parity of reasoning, a desire that we can rationally criticize. This form of rational criticism may be defeasible. For example, we may not rationally criticize a desire that fails to realize its content if it was outweighed by stronger desires. But, in the absence of appropriate defeaters, the criticism stands. If this is right, then it might be thought that an agent who desires to desire to ϕ but who does not desire to ϕ may be rationally criticized after all. Thus (C1).

Does this provide an adequate defence of (C1)? Even if it captures the letter of (C1), it does not capture its spirit. For our earlier discussions suggest that values rationally constrain desires in the following stronger sense. Suppose someone values ϕing, but desires not to ϕ. Such an agent rationally should get rid of the desire not to ϕ and acquire the desire to ϕ *instead*. However, for all we have said so far, he should merely acquire the desire to ϕ *as well*. Can we derive the stronger conclusion that he should get rid of his desire not to ϕ from the reduction of valuing to desiring to desire, plus further plausible assumptions? At this point we need to complicate our discussion. We need to ask what further principles rationally constrain desire formation and retention.

So far we have seen how an agent who desires to desire to ϕ, and who desires not to ϕ, may find himself rationally con-

strained to end up desiring to ϕ as well. But now consider his new set of desires. Is he irrational for having this combination? He may certainly seem to be. But why?

The obvious suggestion is that he is irrational because his desires are not co-satisfiable.[20] However, unco-satisfiability all by itself doesn't seem to be enough. For suppose I desire to be a musician and desire to be a philosopher. I do not seem to be irrational if I do not get rid of one desire or the other once I realize that I cannot satisfy both these desires in the world as it is. I may well end up being disappointed, but that is quite another matter.

It might be thought that this response helps focus on what is at issue. The crucial difference between this case and the other is that in this case it is at least logically possible for my desires to be co-satisfied. Where desiring to be both a musician and a philosopher maps out a coherent life, albeit one that is not empirically realizable, desiring both to ϕ and not to ϕ maps out nothing whatsoever. Co-satisfiability *simpliciter* may not be a constraint on rational desire formation and retention, but, the objector might say, the *logical possibility* of co-satisfiability does seem to be such a constraint.

Let's grant this line of thought for the sake of argument. Does it help? With the logical possibility of co-satisfiability under our belt can we show that the reduction of valuing to desiring to desire yields the stronger conclusion that someone who values ϕing, but who also desires not to ϕ, rationally should get rid of the desire not to ϕ and acquire the desire to ϕ instead?

Certainly we can now see that someone who desires to desire to ϕ, and who therefore ends up desiring to ϕ, and yet who desires not to ϕ as well, ends up having a set of desires that is irrational. However, and importantly, nothing said so far tells us why the rational thing for such an agent to do is to give up desiring not to ϕ, as opposed, say, to giving up his desire to ϕ, and, perhaps, his desire to desire to ϕ as well.[21]

[20] I do not want to endorse this suggestion myself, for I am unsure whether desires, as opposed to intentions, need to satisfy any condition of co-satisfiability at all. My concern is rather to show that *even if* desires must satisfy some such condition, it will not help.

[21] Note that the logical possibility of co-satisfiability does not tell us that it is

The logical possibility of co-satisfiability simply tells us that *some* change has to be made that brings about the logical possibility of co-satisfiability. It doesn't tell us that one of his desires is to be rationally preferred to the other, and thus it certainly doesn't tell us which of his desires is to be rationally preferred to the other.

Let's therefore forget co-satisfiability. There is another option. For in order to make the reduction of valuing to desiring to desire consistent with the claim that an agent rationally should desire in accordance with the normative reason claims he accepts, we could simply add to the reduction the following principle:

(D) If an agent desires to desire to ϕ then he rationally should desire to ϕ, and if he desires to desire to ϕ and desires not to ϕ then he rationally should get rid of the desire not to ϕ and acquire the desire to ϕ instead.

But this response misses the point.

(C1) is supposed to act as a constraint on an adequate account of normative reasons. The hope is that, if we can reduce accepting a normative reason to desiring to desire, then the reduction, in conjunction with other plausible assumptions, will actually entail this principle. But (D) is hardly an additional plausible assumption. It is a theoretically motivated principle that we should accept only if we are given an adequate argument. But the only argument we have been given is that it must be true if the reduction of valuing to desiring to desire is correct. And, in this context, that is not a good argument.

Another way of putting the same point is this. The reduction of valuing to desiring to desire in conjunction with (D) does indeed capture the spirit of (C1). But the reduction itself plays no significant role in this. The spirit of (C1) is captured by the conjunction of even the most implausible reduction with a

irrational to have the desire to desire to ϕ and the desire not to ϕ. For these desires *are* co-satisfiable. If having this pair of desires is irrational then we need a further principle to explain why, presumably: 'It is irrational to have a pair of desires if their co-satisfaction brings about the having of a pair of desires that it is logically impossible to co-satisfy.' Thus the added qualification that the agent may have to rid himself of his desire to desire to ϕ as well.

principle, like (D), that stipulates the very connection we want to derive. Things would, of course, be different if we had some *independent* reason to accept the reduction. The point is just that, in the absence of such reasons, we have no reason to prop up the reduction by accepting (D). Let's therefore consider Lewis's reduction on its own terms to see whether any independent reasons are forthcoming.

Those who seek to reduce valuing to higher-order desiring face a formidable objection.[22] Since they seek to reduce valuing to higher-order desiring, they must come clean and identify valuing with higher-order desiring *at some particular level or other*. Lewis does come clean in this way. He identifies valuing with second-order desiring. But now the quite general problem such theorists face can be put in the form of a question for Lewis in particular. Why identify valuing with second-order desiring? Why not third-order, or fourth-order, or . . . ?

The implication of the question is, of course, that each identification is as plausible as any other. But if each is as plausible as any other, then *all* such identifications are equally implausible. Therefore, no plausible reduction has been effected.

Lewis confronts this objection fairly and squarely. However, his response is less than convincing. He tells us that his reason for favouring the second over the first is that 'a thoughtful addict may desire his euphoric daze, but not value it'.[23] And he tells us that his reason for favouring some level other than the highest order at which an agent desires is that 'if we go for the highest order, we automatically rule out the case of someone who desires to value differently than he does, yet this case is not obviously impossible'.[24] So far this line of reasoning seems perfectly sound. However, and unfortunately, there are no more premisses to Lewis's argument. From these premisses he concludes that valuing is second-order desiring.

The problem isn't just that there is a gap between Lewis's

[22] This objection is forcefully developed by Gary Watson in 'Free Agency', 107–9. The target of his attack is Harry Frankfurt in 'Freedom of the Will and the Concept of a Person'. Frankfurt attempts to respond to Watson's criticism, unsuccessfully I think, in 'Identification and Externality', in his *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988).

[23] 'Dispositional Theories of Value', 115.

[24] Ibid. 116.

premisses and his conclusion. The problem is that his argument demonstrates how formidable the original objection which it is supposed to answer really is. For his conclusion is simply arbitrary, given his premisses. He could equally well have chosen any level other than the first or the highest. And that just *is* the original objection.

If this is right then it follows that we cannot identify valuing with desiring to desire *at any level*. This is not to say that someone who values ϕing may not, even perhaps usually, desire to desire to ϕ. It is merely to insist that we not mistake this contingent fact for a conceptual necessity.

8. VALUING AS BELIEVING

If we cannot reduce valuing to desiring then we have no alternative but to consider reducing valuing to believing. Now Lewis, like Davidson, rejects the idea that valuing is believing. But why does he reject this idea?

Lewis reasons as follows:

> *What is valuing?* It is some sort of mental state, directed toward that which is valued. It might be a feeling, or a belief, or a desire. . . . A feeling?—Evidently not, because the feelings we have when we value things are too diverse. A belief? . . . if valuing something just meant having a certain belief about it, then it seems that there would be no conceptual reason why valuing is a favourable attitude. We might not have favoured the things we value. We might have opposed them, or been entirely indifferent. So we turn to desires.[25]

Lewis's argument against identifying valuing with believing thus depends crucially on the idea that there is some sort of conceptual connection between valuing and desiring. But does granting that connection really preclude identifying valuing with believing valuable?

Lewis seems to think it does, but it is not at all clear why. After all, as he himself notes, the addict may desire his euphoric daze, but not value it; and he may value an unaltered, mundane state of consciousness, but not desire it. And to these examples we may add Watson's woman with a

[25] Ibid.

bawling baby; his angry, defeated squash player; and Stocker's depressives. In other words, it isn't just a conceptual possibility, it actually happens that we are indifferent, or opposed, to what we value! Whatever the precise nature of the conceptual connection between valuing and desiring, then, it does not obviously preclude the sort of indifference or opposition to what we value that the identification of valuing with believing valuable makes possible.

What is the nature of the conceptual connection between believing valuable and desiring? The answer supported by the discussion thus far is that the conceptual connection is simply the defeasible connection described in (C1): we *rationally should* desire what we value. If valuing is believing valuable, and believing valuable is believing that we have a normative reason, then the connection we are after is this:

(C2) If an agent believes that he has a normative reason to ϕ, he rationally should desire to ϕ.

And now we have to face the real problem. For how are we to demonstrate the possibility of this kind of conceptual connection between our beliefs and desires?

Plainly our strategy must be to provide an analysis of our concept of a normative reason and then show that that analysis, in conjunction with other plausible assumptions, actually entails (C2). But can we do this?

Certainly there have been attempts. Consider, for example, the following suggestion of Mark Johnston's:

> As for securing an internal or conceptual connection between value and the will, *this* at least is true: to the extent that one is not weak-willed one will desire . . . as one judges valuable. So much is part of the definition of weakness of will. As far as making the connection between judging valuable and desiring . . . particularly intelligible, this seems to me achieved by the observation that 'valuable' and 'desire-worthy' are near synonyms. If judging valuable is pretty much judging desire-worthy then it is readily intelligible why judging valuable should lead to desiring.[26]

This argument might well be convincing if we were to accept the claim to near synonymy. But should we? Are 'valuable' and 'desire-worthy' near synonyms?

[26] Johnston, 'Dispositional Theories of Value', 161.

Certainly, they are not actual synonyms. For whereas if ϕing is valuable then ϕing is worth doing, if ϕing is desire-worthy then ϕing is worth desiring. But 'worth doing' means something different from 'worth desiring'.

However, though not actual synonyms, Johnston may still be right that they are near synonyms. For he might think that it follows from the fact that something is worth doing that it is worth desiring, and that it follows from the fact that something is worth desiring that it is worth doing. But even this seems quite wrong to me. For, as Derek Parfit has recently pointed out, it may not be desirable that we desire to do what it is desirable that we do, and it may not be desirable that we do what it is desirable that we desire to do.[27] Consider Parfit's example.

The self-interest theory tells me that the desirability of an action or a desire is a function of the contribution that that action or desire makes to my long-term self-interest. Thus it is desirable that I *do* just one thing: promote my long-term self-interest. However, as Parfit points out, it does not follow that it is desirable that I desire to promote my long-term self-interest. Indeed, it may well be undesirable that I desire to promote my long-term self-interest. It all depends on whether having that desire is necessary in order for me to have the set of desires the having of which will contribute most to my long-term self-interest. And that desire may well not be necessary, for my long-term self-interest may be best served by my desiring to act for the sake of family and friends, write books, advance humanity, and so on, without having any direct concern whatever for my own long-term self-interest.

If this seems right, then it follows that the self-interest theory tells me that it is desirable that I desire to do what it is not desirable that I do. And it also tells me that it is desirable to do what it is not desirable that I desire to do. However unlikely this may seem, the point to emphasize is that the issue is an *empirical* one, and so requires an empirical answer. The answer is not determined by 'near' conceptual fiat.

The self-interest theory is, of course, just an example. But what is true of the self-interest theory may well be true of other

[27] D. Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), pt. I.

substantive theories of practical rationality, even the correct theory. Johnston's argument overlooks the possibility of this sort of split between what it is desirable that we do and what it is desirable that we desire to do. As such, it fails as an attempt to make 'readily intelligible why judging valuable should lead to desiring'.

However, though Johnston's argument rests on a false claim to near synonymy it is, I think, on exactly the right track. If anything can make intelligible the connection between judging desirable and desiring it is an analysis of our concept of desirability. Let's therefore see whether we can formulate an argument along his lines in a way that avoids Parfit's objection.

We have seen that to say an action is desirable is to say that we have a reason to do it, where the relevant norms of assessment are the norms of rationality. Now note that we can further explicate this concept, the concept of what we have normative reason to do, though in a way rather different from that suggested by Johnston. For it is a platitude to say that *what it is desirable that we do*—that is, what we have reason to do—is *what we would desire to do if we were rational.*

If an argument along Johnston's lines is to be found, then the argument will have to be that this platitude somehow suffices to make readily intelligible why believing desirable should lead to desiring. I believe that we can provide such an argument. However, before giving the argument, let me say a little about the platitude itself in order to forestall some objections.

9. FORESTALLING SOME OBJECTIONS

The platitude tells us that what it is desirable to do is what we would desire to do if we were rational. But is this really a platitude? Is it too vulnerable to Parfit's objection? It might be thought so: 'To say that we would desire to ϕ if we were rational is to say that desiring to ϕ is rationally appropriate. But the claim that *desiring to* ϕ is rationally appropriate is different from the claim that ϕ*ing* is rationally appropriate. Yet what we want is an explication of the latter idea, not the former.'

Suppose, for the sake of argument, that what we have reason to do, in our actual circumstances, is promote our long-

term self-interest. According to the platitude it follows that, if I were rational, I would desire that, when in my actual circumstances, I promote my long-term self-interest. But this is not to say that that desire is rationally appropriate in a way that is vulnerable to Parfit's objection.

In order to see this, note that the platitude concerns a desire I would have, if I were rational, about what I am to *do*, in my actual circumstances. It does not concern a desire I actually have in my actual circumstances. And nor does it concern a desire I would have, if I were rational, about what I am to *desire* in my actual circumstances. Thus, for all the platitude tells us, what it is rational for me to do and what it is rational for me to desire may be quite different. It all depends on whether, if I were rational, what I would desire that I do, in a given set of circumstances, and what I would desire that I desire, in a given set of circumstances, are the same.

Desirability is, then, a function of our rationally appropriate desires. But what is desirable is not the having of those desires themselves, but rather the *objects* of those desires. It thus seems to me quite safe to say that the platitude avoids Parfit's objection.

Even if the platitude avoids Parfit's objection, it might be thought vulnerable to an objection from another direction: 'If the platitude is right, "ϕing is what I would desire if I were rational" gives the content of the thought "ϕing is desirable". Thoughts about our values thus turn out to be thoughts about our desires! But this seems wrong. When we deliberate we do not focus our attention in on ourselves and our own desires, we focus our attention out on the value of what we desire. Value judgements are not introspective claims about our desires, they are claims about a standard against which our desires are measured.' This objection is wrong in two ways.

First, the platitude does not tell us that thoughts about our values focus in on ourselves and our own desires. Instead, it offers us a striking contrast between introspective judgements about our own desires and our value judgements. In order to see this, contrast two ways in which a desire may figure in deliberation if, as the objector supposes, the platitude gives the content of our thoughts about our values (I question this assumption presently).

Suppose *A* very much wants to dance a jig. His desire may be taken into account in his deliberations in the following way. Recognizing his desire, he considers it and decides that it is a desire both worth having and worth acting upon. It will be fun to dance a jig, for it is fun to give one's body over to the music and move in the regular pattern that it dictates.

When *A* deliberates in this way, how does desire figure in his decision-making? Desire figures in his decision-making in at least two ways. First, introspected, his desire to dance a jig figures as an object of positive evaluation. This is indeed a case in which *A* focuses in on himself and his own desires. However, that is hardly surprising given that it is, *inter alia*, his desire to dance a jig that is being evaluated. Secondly, however, the fact that *A* would desire that, in his actual circumstances, he both desires to dance a jig and acts on that desire also figures in his decision-making. But is this an introspective judgement? Certainly not. For what makes the judgement true is not an introspectible fact about *A*; rather, it is a hypothetical fact about *A*: that is, a fact about what he would want if he were rational.

Moreover, whereas the introspective claim about *A* that he has a certain desire, say the desire to dance a jig, gives us no reason to suppose that his desire is worth having or worth acting upon, the fact that, if he were rational, *A* would want that, in his actual circumstances, he both has that desire and acts upon it does give us such a reason. This hypothetical fact about *A*'s desire, then, does seem to offer us a standard against which his introspectible desires may be measured, just as the objector insists.

Even if the platitude does give the content of our evaluative thoughts, then, our evaluative thoughts are not thereby made introspective judgements. They are introspective only if the *object* of evaluation is an introspectible item. Moreover, as we have just seen, the platitude does suffice to show how values constitute a standard against which our introspectible desires may be measured. It thus seems to me that the objection simply misses the mark.[28]

[28] There is, in fact, a third way in which desire may figure in decision-making. Suppose *B* has had a strict puritanical upbringing. He has been taught, and he believes, that bodily pleasures are to be avoided, that they

I said that the objection is wrong in two ways. The second is in the assumption that the platitude constitutes, or holds the place for, a *reductive analysis* of our concept of desirability. This assumption is implicit in the claim that the platitude gives the content of our thoughts about desirability. But quite the opposite is in fact true. Though the platitude can be refined, it cannot be turned into a reductive analysis. Thus, contrary to the objection, the platitude does not even entail that evaluative thoughts are thoughts about our own hypothetical desires.

The idea that the platitude holds the place for a reductive analysis of our concept of desirability, or a normative reason, is a common one. The idea seems to be that we can turn the platitude into a reductive analysis by giving a substantive account of what is required in order to be 'rational'. Perhaps the best-known recent attempt is Bernard Williams's in his 'Internal and External Reasons'.[29] He offers us the following:

A has a reason to ϕ in circumstances *C* if and only if *A* would desire that he ϕ's, in circumstances *C*, if:

(1) *A* had no false beliefs,
(2) *A* had all relevant true beliefs,
(3) *A* deliberated correctly.

Williams motivates the claim that these conditions constrain what counts as a reason by focusing on particular examples.

corrupt. Moreover, suppose he believes, plausibly, that dancing a jig is a way of getting bodily pleasure. And suppose further that *B* finds himself wanting desperately to dance a jig. Imagine how his desire would be taken into account in his deliberations. *B*'s desire certainly doesn't figure in his deliberations as the object of a positive evaluation. Indeed, if anything, it figures in his deliberations as the object of a negative evaluation. In deciding what he can justify doing, *B* may well have to conclude, therefore, that he cannot justify acting on his desire to dance a jig. Indeed, he may have to conclude that he cannot even justify having the desire. This may be the conclusion to which his deliberations lead him. However, the fact remains that *B* has the desire. And, alienated from it though he may be, he may find it simply irresistible. Now, it seems, we have a desire figuring in decision-making quite *independently* of deliberation. For in this case *B*'s recognition of the fact that he has the desire seems to determine his decision about what he will do *independently* of his judgements of value. This case is, of course, only one of many. The Ayer–Frankfurt–Watson cases all suggest that desires may figure in decision-making in something like this way. Philip Pettit and I mention cases of this kind in our discussion of capriciousness in 'Backgrounding Desire', §3.4. We hope to discuss these cases in greater detail elsewhere.

[29] B. Williams, 'Internal and External Reasons', in B. Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981).

Consider (1). Suppose an agent desires to mix some stuff from a certain bottle with tonic and drink it. However, he has this desire only because he desires to drink a gin and tonic and believes that the bottle contains gin, whereas in fact the bottle contains petrol. As Williams points out: 'it is just very odd to say that he has a reason to drink this stuff, and natural to say that he has no reason to drink it, although he thinks that he has'.[30] Why? Because he would not have the desire if he were rational: that is, *if he had no false beliefs.*

Consider (2). Suppose I desire to buy a Picasso. Moreover, suppose that, without my knowing, there is a Picasso for sale very cheap in the local second-hand shop. It would certainly be both true and appropriate for a friend to tell me that I have a reason to buy something from that shop. For, quite in general, as Williams says, an agent 'may be ignorant of some fact such that if he did know it he would, in virtue of some element in [his set of desires] . . . be disposed to ϕ: we can say that he has a reason to ϕ, though he does not know it'.[31] Why? Because he would desire to ϕ if he were rational: that is, *if he had all relevant true beliefs.*

Now consider (3). So far we have taken it for granted that desires and beliefs interact in ways that generate new desires. But this is, of course, a substantive claim about practical reason. Our desires and beliefs only generate new desires if we deliberate correctly: that is, *inter alia*, according to the means–ends principle. 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 ϕing 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: eg. by time-ordering; where there is some irresoluble conflict among the elements of [one's set of desires] . . .

[30] Ibid. 102.
[31] Ibid. 103.

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.[32]

I will return to this point presently. For now, simply note that, given the wide variety of principles that govern rational deliberation, an agent has a reason to ϕ only if he would desire to if, in addition to the other constraints, his deliberations conform to these principles: that is, *if he deliberates correctly*.

So far so good. However, Williams doesn't offer us any more constraints on what is to count as a reason. But, as should be evident given the concerns of this paper, more constraints are certainly needed if we are to succeed in giving a reduction.

Consider, for example, Watson's woman who desires to drown her bawling baby in the bathwater. As we saw earlier, her desires are, even by her own lights, *unreasonable*. She desires to do what she believes she has no reason to do. But Williams's reduction does not have this conclusion. Her desire is not the result of any false belief; it would not go away if she were to acquire some further true belief; and nor does she seem to be suffering from some deliberative failure, or at least, not any failure of the kind Williams describes. And, of course, there are other examples as well: Watson's defeated squash player, not to mention Stocker's depressives.

It might be thought that these examples simply require the addition of a further condition:

(4) *A* is in a normal emotional state.

Nor, it might be said, is this *ad hoc*. It merely assumes, rightly, that we give rational privilege to the desires we would have if we were in such a state.

The additional condition certainly succeeds in telling us that Watson's woman with a bawling baby has no reason to drown her baby. For though she actually desires to drown her baby, if she were in a normal emotional state, she would not desire that, when suffering severe emotional distress due to the bawling of her baby, she actually drowns it. And perhaps the

[32] Ibid. 104.

additional condition would let us deal with Watson's angry, defeated squash player and Stocker's depressives too.

However, though not at all *ad hoc*, and even if successful at dealing with these examples, it seems to me that the need for a condition like (4) signals the end of the search for a reductive analysis. For the analysis is truly *reductive* only if, in specifying what is to count as a 'normal' emotional state, we need make no reference to what we have reason to do. But I see no reason to suppose that this could be done. For we have no grip on what is to count as a 'normal' emotional state except in the context of our practice of giving reasons and excusing failures.[33] That this is so is manifest in the fact that we simply add to our list of what is to count as a 'normal' emotional state as further examples crop up. What guides us is our conception of what is to count as a good reason or an excuse, not the intrinsic nature of the emotional state itself.

And matters are, of course, in fact much worse. For other examples force further revisions. If the heroin addict's and the kleptomaniac's desires do not provide them with reasons, and if we have no reason to suppose that their desires would disappear if they satisfied conditions (1) to (4), then we need to add a further condition, presumably:

(5) *A* is in a normal physical state.

And, once again, though not at all *ad hoc*, and even if successful at dealing with these examples, the need for a condition like (5) signals the end of the search for a reductive analysis. For, again, we have no grip on what is to count as a 'normal' physical state outside of our practice of giving reasons and excusing failures.

It is important to note what this implies. It does not imply that the platitude is not a platitude. For, equally, just as the platitude tells us, we have no grip on what is to count as a reason except in terms of what we would desire if we were rational. What it implies is rather that this platitude cannot be honed into a reductive analysis. What we have is, if you like, a non-reductive 'explication' of our concept of a reason.

[33] Stocker makes a similar point at the end of 'Desiring the Bad: An Essay in Moral Psychology'.

If this is right then, significantly, a central theme in Williams's work on reasons is shown to be fundamentally flawed. For, in embracing the platitude, Williams takes himself to be defending a *relativized* conception of reasons. As he puts it:

> the 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 ϕing, 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.[34]

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* . . .[35]

But what is the relativity?

Agreeing that an agent has a reason to ϕ only if he would desire to ϕ if he were rational, Williams insists that what an agent would desire if he were rational is relative to the desires he actually has; his rational desires are, as it were, *functions from* his actual desires, where the functions are those described in conditions (1) to (3). Thus, he claims, we cannot expect that, if we were all rational, different agents would converge upon a unique set of desires. This is why Williams claims to be defending a Humean conception of reasons.[36] For he denies that rational agents all have the *same* reasons.[37]

If (1) to (3) constituted a reductive analysis of a reason then it seems to me that we can see why Williams makes this assertion. For the only element in the analysis that holds out the hope for a convergence in the desires of rational agents is condition (3): the requirement that agents deliberate correctly. But if we give this condition a reductive construal, then that hope simply fades away.

Consider, for example, the most revisionary of the forms of deliberation Williams proposes. He tells us that 'where there

[34] Williams, 'Internal and External Reasons', 101.
[35] Ibid. 102.
[36] Ibid.
[37] This comes out most clearly in his discussion of the Owen Wingrave example: ibid. 108–11.

is some irresoluble conflict among the elements' of one's set of desires one must consider 'which one attaches most weight to'. Now this sounds like a familiar overall consistency and coherence requirement. Construed as a function performed on an agent's actual desires, we can imagine that, in the interests of overall consistency and coherence, the agent may lose certain desires and acquire others. But the scope for this sort of revision is, we might think, severely limited. For what the agent is after is a consistent and coherent set of desires, given his actual desires as the starting-point. Thus, once we add in the fact that agents differ wildly in the desires that they actually have, we see no reason to suppose that, even as a matter of *empirical* fact, when this function is performed on the desires of agents with diverse sets of desires, they will end up with the same set of desires. Still less do we see reason to suppose that they would converge on a set of desires as a matter of *rational* fact. In other words, rational agents may yet differ in the desires that they have.

However, note how different things look when we construe condition (3) non-reductively. For the starting-point of deliberation is not now the agent's actual desires, but the value judgements he actually believes. And, according to condition (3), what he is required to do is to revise them so as to make for overall consistency and coherence.

Here, again, there is scope for the agent to give up some of his values, on the grounds that they are false, and acquire new values, on the grounds that they allow him to make better sense of the rest of his values. But it seems that, in this case, the scope for massive revisions in an agent's evaluative beliefs is much greater. Equip the agent with a robust sense of his own fallibility and a disposition to consider the evaluative beliefs of others as in conflict with his own, as opposed to merely different from his own, and we are well on the way to the possibility of a convergence, a convergence mandated by reason itself.[38]

Now there is certainly no proof that rational agents would actually end up converging on a single set of values. But,

[38] Myself I think that this is one way of reading Rawls's argument in his 'Outline of a Decision Procedure for Ethics', *Philosophical Review*, 60 (1951).

equally, there is no proof that they would not. There is simply no way of telling in advance. We must give the justifications and see where the arguments lead. But, if this is right, then it is simply a bald assertion to claim, as Williams does, that the platitude itself implies that our reasons are *relative*.[39]

10. THE PUZZLE SOLVED

I said the platitude that desirability is simply a matter of what we would desire if we were rational suffices for making sense of (C2): the claim that if I believe that I have a normative reason to ϕ, then I rationally should desire to ϕ. Let me now explain why that is so.

(C2) tells us that if we believe we have a normative reason to ϕ then we rationally should desire to ϕ. And that is surely no surprise if believing that we have a normative reason to ϕ, or that it is desirable that we ϕ, amounts to the belief that we would desire to ϕ if we were rational. For suppose we believe that we would desire to ϕ if we were rational and yet fail to desire to ϕ. Are we irrational? We most certainly are. And by our own lights! For we fail to have a desire that we believe it is rational for us to have. In other words, if we believe that we

[39] More formally the point can be put like this. Suppose that the proper analysis of the claim 'It is desirable that p' is relative. Thus: 'It is desirable$_x$ that p iff x would desire that p if x were rational', where 'desirable$_x$' is to be read as 'desirable-from-x's-point-of-view'. The relativization is not idle so long as a convergence in judgements of desirability does not emerge under conditions of full rationality. But if such a convergence were to emerge, the relativization would become idle; for then everyone's point of view would be the same as everyone else's. And now the point can be put like this: we can discover whether relative reasons are *the only reasons that there are* by seeing whether the arguments for and against our judgements of desirability lead us to converge on a single set of judgements of desirability. (Note that I have here abstracted away from the purely conceptual question 'Do we think of our judgements of desirability as relative (that is, as implicitly indexed to individuals) or as non-relative (that is, as implicitly universally quantified)?' My own view is that we think of our judgements of desirability as non-relative: that is, that the proper analysis of our judgements of desirability would have them concern what *we* would want if we were fully rational. Only so can we explain why we think of ourselves as in disagreement with, as opposed to merely different from, each other, in so far as we accept different evaluations. For more on this see my 'Dispositional Theories of Value' and 'Realism'.)

would desire to ϕ if we were rational then we rationally should desire to ϕ. And that is just (C2).

In this way we capture the letter of (C2), but can we capture its spirit? If we believe that we would desire to ϕ if we were rational, and yet desire not to ϕ, can we see why we should get rid of the desire not to ϕ and acquire the desire to ϕ instead, rather than, for example, change our evaluative belief? (Here we recall the problem facing the reduction of valuing to desiring to desire.)

We certainly can. Remember, ϕing is desirable just in case we would desire to ϕ if we were rational. Now, by hypothesis, what we believe is that we would desire to ϕ if we were rational. We do not believe that we would desire *not* to ϕ if we were rational. And the mere fact that we actually desire not to ϕ gives us no reason to change this belief. Believing what we believe, it therefore follows that we rationally should get rid of the desire not to ϕ and acquire the desire to ϕ instead.[40]

This argument is admittedly very simple. As with many simple arguments, its real power may therefore be overlooked; it might be thought 'too simple'. So let me add further support for this argument by showing that a structurally similar argument allows us to explain a similar phenomenon in the case of belief.

[40] Is this argument consistent with the earlier argument against Johnston? It is. Indeed it helps to explain why a theory like the self-interest theory may be self-effacing. Suppose it is rational for me to do just one thing: promote my long-term self-interest. And suppose further that it is not rational for me to desire to promote my long-term self-interest; that my long-term self-interest would be best served by my desiring to act for the sake of family and friends, write books, advance humanity, and so on, without having any direct concern whatever for my own long-term self-interest. What it is desirable that I do is, in this case, not what it is desirable that I desire that I do. But now suppose I come to *believe* the self-interest theory. I come to believe that it is desirable to promote my long-term self-interest and undesirable to desire to promote my long-term self-interest. From the argument just given, having these beliefs makes it rational for me to desire to promote my long-term self-interest and to desire not to desire to promote my own long-term self-interest. Since the reason I have the desire to promote my long-term self-interest, something we know independently that I rationally shouldn't desire, is that I *believe* the self-interest theory, it is no surprise to learn that I rationally shouldn't believe the self-interest theory. The theory is self-effacing. And since I desire not to desire to promote my long-term self-interest, it is no surprise that I am motivated to get rid of that belief. I am indeed moved to do what the theory tells me it is rational to do.

Note that the following principle, itself much like (C2), governs our beliefs:

(C3) If an agent believes he has (most) reason to believe that p then he rationally should believe that p.

And note, furthermore, that we can explain (C3) via an argument that strictly parallels the argument just given to explain (C2).

That argument trades on a platitude about reasons for action. So consider a platitude about reasons for believing. Just as, if we have a reason for ϕing, we can say that ϕing is 'desirable', where desirability is fixed by norms of rationality, if we have (most) reason to believe that p, we can say that p is (most) 'believable', where believability is fixed by norms of rationality. But now note that just as it is a platitude to say that if ϕing is desirable then ϕing is what we would desire if we were rational, it is also a platitude to say that if p is (most) believable then p is what we would believe if we were rational. Equipped with these platitudes, we have enough to explain (C3).

Suppose an agent believes that he would believe that p if he were rational and yet fails to believe that p. Is he irrational? He certainly is. And by his own lights! For he fails to believe something he believes he has good reason to believe. Indeed, this must surely be a paradigmatic case of irrationality.

Moreover, note that we can also explain why someone who believes that p is (most) believable, but who also finds himself believing that not-p, rationally should get rid of his belief that not-p and acquire the belief that p instead. For p is (most) believable just in case he would believe that p if he were rational. And, by hypothesis, that is what he believes. He does not believe that he would believe that not-p if he were rational. And the mere fact that he actually believes that not-p gives him no reason to change this belief. Thus he rationally should get rid of his belief that not-p and acquire the belief that p instead. And that is just (C3).

Given the structural similarity between this argument and the argument for (C2), and given the success of the argument in the case of belief, I conclude that both arguments are successful. The platitude that desirability is a matter of what we would desire if we were rational suffices to show how it can

be that our beliefs about our reasons rationally require us to have corresponding desires.

11. RECONCILIATION

It seemed difficult to reconcile the claim that deliberation on the basis of our values is practical in its issue to just the extent that it is with two further claims, the claim that deliberation normally reflects our evaluative *beliefs* and the claim that our actions are produced by our *desires*. However, we have seen that these claims are not in conflict. Instead they simply reflect a substantive fact about human agents: namely, that we are rational creatures who are sometimes more rational, sometimes less. Deliberation on the basis of our evaluative beliefs is practical in its issue *to just the extent that it is* because *that* is precisely the extent to which we are rational.

The point is not that this answer is in any way surprising. It was always the only answer available. The point is rather that now we know *why* this is the answer—what the *role* of our being rational is. For if, when we deliberate, we take into account what we have reason to do, and what we have reason to do is a matter of what we would desire to do if we were rational, and to the extent that we are rational we will desire to do what we believe we would desire to do if we were rational, then nothing else but the contingent fact that we are rational to just the extent that we are *could* explain why deliberation on the basis of our values is practical in its issue to just the extent that it is. Our contingent rationality is the only variable. The contingent fact that we are rational thus matches our motivating reasons with our beliefs about our normative reasons.

Given this reconciliation, it follows that there is no conflict in the two perspectives on the explanation of action described at the outset: the intentional and the deliberative. All intentional actions are indeed explicable from the intentional perspective in terms of our underlying desires and beliefs. But, to the extent that we are rational, our actions are also explicable from the deliberative perspective. For when we deliberate we concern ourselves with our normative reasons, and, to the extent that we are rational, our underlying desires will match

our beliefs about the normative reasons that we have. It is thus our substantive rationality that explains why the connection between deliberation and action is not entirely contingent and fortuitous.

Finally, despite the central importance of the platitude linking what is desirable with what we would desire if we were rational in effecting this reconciliation, we have seen that we cannot turn this platitude into a fully reductive account of our concept of desirability. This has important implications for how we conceive of the reasons that we have, for it holds out the possibility of a convergence in our judgements about what it is desirable that we do under conditions of full rationality. The conception of ourselves as deliberating agents described at the outset does not require such a convergence. However, it is fitting, given the central importance of that conception of ourselves in Kantian circles, that in defending it, we come to see a non-relative account of our reasons for action as a real possibility.