OXFORD JOURNALS
OXFORD UNIVERSITY PRESS

Mind Association

Practical Unreason
Author(s): Philip Pettit and Michael Smith
Source: *Mind*, New Series, Vol. 102, No. 405 (Jan., 1993), pp. 53-79
Published by: Oxford University Press on behalf of the Mind Association
Stable URL: http://www.jstor.org/stable/2254172
Accessed: 27/10/2008 09:25

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=oup.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

Oxford University Press and *Mind Association* are collaborating with JSTOR to digitize, preserve and extend access to *Mind.*

http://www.jstor.org

Practical Unreason

PHILIP PETTIT and MICHAEL SMITH

The philosophical literature on failures of practical reason generally takes categories of failure recognised in common-sense morality and in the philosophical tradition—weakness of will, compulsion, wantonness and the like—and offers a reconstruction of what is involved in such failures. The approach is deferential; it casts philosophy in the role of underlabourer to received wisdom. In this paper we explore a methodologically bolder approach to practical irrationality. We start with a distinction between intentional and deliberative perspectives on the explanation of action and we try to show how it can be used to generate a systematic taxonomy of the different types of failure that we may expect to find in practical reason.

The approach which we explore is not only methodologically bolder than the standard approach; it also differs substantively. Some contemporary theories treat phenomena like weakness of will, compulsion and wantonness as practical failures but not as failures of rationality: say, as failures of autonomy or whatever. Other current theories—the majority—see the phenomena as failures of rationality but not as distinctively practical failures. They depict them as always involving a theoretical deficiency: a sort of ignorance, error, inattention or illogic. They represent them as failures which are on a par with breakdowns of theoretical reason; the failures may not have exact theoretical analogues, exact analogues in the breakdown of belief, but they are of essentially the same, cognitive kind. Our approach gives us quite a different view of things. The pathologies which we identify in our taxonomy are distinctively rational failures and distinctively practical failures; they are failures of pure practical reason.

The paper is in five main sections. In section one we introduce the distinction between the intentional and the deliberative dimensions of decision-making and in section two we describe an ideal of practical rationality in which the intentional dimension is in resonance with the deliberative. This puts us in a position, in the third section, to look at the ways in which that resonance can break down and at the corresponding failures of practical reason; the breakdown of resonance may lead to outright dissonance, as we describe it, or to a mere consonance between the two dimensions. The last two sections provide a commentary on the position developed in this discussion. In section four we elaborate a little on the methodological and substantive features that mark it off from more standard approaches. And in a final, concluding section we characterise our approach as one under which the heteronomy characteristic of practical unreason contrasts, not with self-rule or autonomy, but with right rule or "orthonomy".

Mind, Vol. 102 . 405 . January 1993 **© Oxford University Press 1993**

1. The intentional and the deliberative

Human beings, we assume, are deliberative agents. As they face a choice, they are capable of registering considerations relevant, by their own lights, to what should be done: thus they can register that these are the alternative options and those the associated possible outcomes, that one option has this set of desirable features, another a different set, and so on. They are capable, furthermore, of registering that the considerations overall support one or another choice: they can recognise the import of the desirable features registered. And they are capable, finally, of being moved by such a pattern of reasoning: they are capable of making this or that choice in response to the recognition that it is the most strongly supported alternative.

We believe that human agents exercise their deliberative capacity to a limited extent in almost every choice—this exercise, as we shall see, may be successful or unsuccessful—but in any case we shall be concerned only with choices where there is deliberation. That agents regularly deliberate does not mean that they explicitly weigh the pros and cons relevant to every choice. We think that in approaching action human agents register the presence and the import of properties that argue for one or another choice; that is why it is reasonable to ask an agent why she thought her action desirable or to ask how she could have been indifferent to features that made it clearly undesirable. But the deliberative registering of the presence and import of such properties may be a subliminal process that is difficult to reconstruct afterwards. Moreover, the process is usually going to be a very incomplete train of reflection; it is going to direct the agent to some properties relevant for the choice but almost certainly not to all.

Where our first assumption is that human beings are deliberative agents, our second is that they are also intentional subjects. The main element in this assumption is the assertion that when humans are moved by deliberative reasoning, not only do their beliefs about the desirability of the features registered play a role in generating action, there is also a role to be played by desires. Beliefs alone are not sufficient for the production of behaviour (Smith 1987).

How can the intentional conception of agents be squared with the deliberative? How can we find a role for desire in those cases where an agent registers that one option has certain desirable properties that are unregistered in alternatives; where she registers that that option is therefore the most desirable; and where she is moved to action by that reasoning?

The straightforward answer to that question is that in such a case, the agent must desire to realise the properties deemed desirable—she must prize or value those properties, at least in the circumstances on hand—and she must desire their realisation with sufficient strength for this to lead to a desire to perform the option that bears them: to perform that option rather than any alternative. The idea is that as the agent registers the considerations relevant in deliberation, not only does she form appropriate beliefs in the presence and import of the properties registered; she also forms desires for the realisation of those properties and ultimately,

as a net effect of such desires, she conceives a desire to perform the appropriate action.

This answer, we say, is straightforward. Our attitude may reflect a third assumption we make, apart from the assumptions associated with the deliberative and intentional conceptions. We assume that when a human agent comes to form a desire for this or that option among the alternatives that face her in a decision, she does so as a result of desiring to realise certain properties that she expects the option or its outcome to instantiate.[1] In particular, she does so as a result of having a stronger desire to realise those properties than any desire she may have vis-à-vis the properties associated with other options. If the activation of property-desires generates option-desire in this way, then it is natural to think that when deliberation issues in action, the agent forms desires for the deliberatively favoured properties and these are sufficient to produce the deliberatively supported action.

Our picture of how the deliberative and intentional conceptions go together—our picture of how deliberation and desire fit with one another—raises an obvious question. Do agents always act as they deem to be most desirable in deliberation? Do their desires always answer to their desirability-beliefs? We conjecture that they do not and that this is what creates an opening for a distinctive form of practical unreason. We hold that an agent may deliberatively favour one option without this impacting suitably on what she desires and what she does. She may choose a different option or she may choose the favoured option but not for the reasons it is deliberatively supported.

Most of our paper amounts to an elaboration of this conjecture and we hope that that elaboration, with examples, will make the conjecture plausible. But we think that it should be more or less obvious, in any case, that people can form desires that diverge from what they believe desirable. A heroin addict may think that there is nothing at all to be said for jabbing the needle into her veins; she may resent the "high" that it gives her and may wish to be rid of the desire for heroin. Yet she may give herself the injection, and do so intentionally, none the less (Frankfurt 1988). A woman may know full well that there is nothing at all to be said for drowning her baby in the bathwater, no consideration that should be outweighed by other reasons. Yet she may do so, out of a sudden whim, and do so intentionally: that is, do so on the usual belief-desire basis (Watson 1982). In cases like these, the action explicable from the intentional perspective does not have properties in virtue of which it presents itself as desirable to the agent, although it may have properties that engage with the agent's desires.

[1] On property-desires see Jackson 1985 and Pettit 1991a: this tries to square property-desires with a decision-theoretic framework. To desire an option is to prefer it to the feasible alternatives. To desire a property is to be disposed, as between options or, more generally, prospects that otherwise leave one indifferent, to prefer a prospect with the property to any prospects without. Thus, if the actual world lacks the property, it is to prefer that it should have the property: it is to prefer the counterfactual world in which the property is realised, assuming that its realisation leaves other things equal.

Such examples show that the intentional and deliberative dimensions of decision-making may indeed come apart.[2] In particular, they show that the conclusion that a certain action is or appears more desirable than alternatives may or may not go hand in hand with an agent's desiring it. A person may conclude that a particular option is desirable, yet not desire it; and a person may desire a particular option, yet not believe it desirable. The divergence between the intentional and the deliberative dimensions of decision-making is not surprising, on our preferred account of the concept of desirability. We take it that an action is desirable in certain circumstances just in case, if the agent were fully rational, she would desire that, were she in those circumstances, she performs an action of that kind (Smith 1992, Pettit and Smith forthcoming).[3] Given this analysis of the concept of desirability it is certainly possible for an agent to come to believe a certain action to be desirable and yet not desire to act in that way, and it is equally possible for her to desire to act in a certain way but not believe that acting in that way is desirable. And so we have an explanation of why the intentional and the deliberative perspectives may come apart in the way that they do. Moreover, given this analysis, we must also suppose that, other things being equal, an agent manifests a form of unreason in not desiring to act in the way she believes desirable. For, by her own lights, she fails to desire to act in the way she would desire to act if she were fully rational. She is therefore irrational by her own lights. And so we have an explanation of why, in agents who are in this respect rational, the two perspectives march in step.

Our purpose in this paper, however, is not to defend this particular account of desirability, nor to address other problems related to how the intentional and the deliberative dimensions of decision-making can come apart. Putting those issues aside, we conjecture that an agent's desires can come apart from her deliberative judgments and our aim is to show how that hypothesis facilitates the characterisation of practical unreason.[4]

[2] We have addressed elsewhere some of the problems generated by the relationship between the intentional and the deliberative dimensions. See Pettit and Smith 1990, forthcoming; Pettit 1991a; Smith 1992.

[3] For a characterisation of this "response-dependent" style of accounting for concepts see Johnston 1989, Pettit 1991b.

[4] Among the issues we would like to put aside is the question of whether desire for something is always or ever necessitated just by the belief that that thing is desirable: whether desire can be a cognitive state. We write in a way that may favour non-cognitivism, arguing that a failure of reason, in particular a failure unparalleled in the theoretical forum, can cause a divergence between desirability-beliefs and desires. But the cognitivist can give a congenial reading to the claim. The weak cognitivist will have no problem in doing so: she thinks that while a desirability-belief necessitates the presence of a corresponding desire, it does not determine the degree of strength of that desire, and so she can regard it as a triumph of practical reason that an agent forms a desire of the appropriate strength. The strong cognitivist, on the face of it, will face a problem. She goes beyond the weak position and holds that the necessitation of desire extends to its degree of strength. She will have to say that if divergence occurs, then the desirability-belief is not held on the proper basis or internalised in the proper way or something of the kind. Thus she will have

2. *Practical rationality*

Before looking at how the intentional and deliberative dimensions can diverge, it will be useful to examine what happens when they converge. Before looking at the different modes of practical irrationality, it will be useful to examine what practical rationality involves. We shall give a sketch of what it is for a particular action to be rational. And then we shall add some details about what is required for an agent, as distinct from an action, to display rationality.

Our discussion of the intentional and deliberative dimensions of decision-making already gives us a picture of how a rational choice will be made, and a rational action produced. The agent will register different desirability-relevant properties in the options, and in the likely outcomes of the options, before her: different values which the options would instantiate or would be likely to instantiate; she will register, for example, that returning this book would fulfil a promise, not returning it would break one. The values registered will be properties in the light of which she tends to desire options, properties which she cherishes or prizes. At a certain point, the properties considered—together, of course, with any relevant outcome-probabilities—will lead the agent to see one particular option, say returning the book, as imperative or prescriptive: as the thing for her to do. And this deliberative judgment, this more or less explicit self-prescription as to what she ought now to do, will be matched by a suitable desire: a desire for the apparently prescriptive option. As the valued properties combine to support the deliberative judgment, so they will combine to produce a desire for the option that is deliberatively favoured.

The dual aspect, deliberative and intentional, of the properties registered in decision-making is the key to this picture of rational action. In rational action the values which lead an agent to prescribe one option to herself—to see it as desirable, all things considered—are also the values which lead her to choose that option. The values that weigh with the agent in deliberation serve also to arouse a desire for the option which they deliberatively support. Their net impact in arousing desire—their net desiderative force—corresponds to their net deliberative weight. As the agent deliberates, so does she desire.

This picture of rational action is drawn briskly, as the details need not concern us, but there are a number of points we should notice.

a. We refer to the perception of an option as prescriptive or as desirable all things considered. This is the perception or judgment which the agent forms, having considered all things—or at least having considered all things that strike her, in the circumstances, as relevant; it is the agent's

---

to acknowledge that the divergence involves a cognitive dimension and displays a certain parallel with failures of theoretical reason. But the strong cognitivist will still be able to identify distinctive—in particular, distinctively practical—features in the failure described: the failure will not amount to any familiar kind of ignorance or error, inattention or illogic. And so she too can endorse the claim defended here.

final or operative judgment of desirability. Notice that the final judgment in this sense is distinct from the judgment that an option is desirable relative-to-all-considerations: that it is, as we might put it, desirable-all-things-considered. (Davidson 1980)

b. In the example of returning the book, the property which weighs with the agent is one that the choice of that option is bound to satisfy: that of keeping a promise. We shall generally speak, for simplicity, as if the properties that register with an agent in producing a choice are properties like this, which are certain of realisation by the appropriate option. But it should be remembered that in most cases the properties that register in decision-making will be just probabilistically connected with the relevant option; they will be properties of outcomes which the option has only a certain probability of bringing about.

c. We only mention the valued properties of options, ignoring their disvalued counterparts. This is legitimate, as disvalued or costly properties in any option can be represented as values or benefits of the alternatives. That any option has a given cost means that the alternatives confer the benefit of avoiding that cost.

d. The valued properties of an option will make it seem prescriptive or desirable, only given the weights which the agent attaches to the values in her reasoning. We allow that the weights attached to values may be indeterminate, so that the judgment of desirability can often be underdetermined. And we allow that the weights attached to certain values may differ between different agents. But we shall not be commenting explicitly on those possibilities.

So much for our picture of the rational action. What of the rational agent? The rational agent will certainly produce rational actions. But she must not produce them as a matter of good luck; she must be someone who produces rational actions reliably. So what is going to be required for a person to be a reliable source of rational actions?

The net deliberative weight of a set of values—the net support it gives to the option prescribed—is determined by the different weights associated with each of those values. And the net desiderative force of a set of values—its net impact in producing desire—is determined by the forces associated with the desires for those values. If an agent is to be reliably rational in the choices she makes, if the net desiderative force is reliably to correspond to net deliberative weight, then two conditions must be fulfilled.

First, all the desiderative forces that determine what an agent does must be determined by the deliberative weights of values. There must be no desires of the kind that move the heroin addict and the distressed parent; there must be no desires that are formed without regard to values. And, second, the desiderative forces contributed by the different values registered must correspond suitably to their deliberative weights. The weight which a value is ascribed in the balance of deliberation must fix the force which the value exercises in the generation of desire; it must determine the strength with which the agent desires to choose an option with that property. If the agent attaches a certain value to helping her friends, for example, then the strength of the corresponding desire should not be

so low that a consideration to which she gives lesser deliberative weight can sway her in a different direction; and it should not be so high that a consideration to which she gives greater weight is unable to deflect her inclination.

Our image of the rational agent, then, is one of a person in balance: a person in whom desiderative forces are matched to deliberative weights. The language of weights and forces is metaphorical but it is not empty.[5] That a value has a certain deliberative weight means, more prosaically, that the agent is disposed to give it a certain importance vis-à-vis other values. That a value has a certain desiderative force means that the recognition of the presence of the property valued generates a desire with a certain strength: with a certain capacity to win out over the desires occasioned by other evaluations or occasioned exogenously. Where the measure of deliberative weight is given by the agent's reasoning practices, the measure of desiderative force is given by her dispositions to action.

This completes what we need to say about practical rationality. It remains only to comment on an objection. Our account may be resisted on the ground that even if an agent desires as she deliberates, and even if she does this reliably, the action which she produces on a given occasion may be irrational in other ways. An action is irrational, it appears, if the deliberative prescription is not actually supported by the valued properties which the agent registers, even if the desire she forms matches that judgment; in this case the agent displays inferential failure. Again, an action is apparently irrational if the valued properties actively registered are not all of those which the agent takes as relevant in, say, earlier reflection, or if they are not weighed as in earlier reflection: the agent displays a selective or biassed attention to the values on offer, being unfaithful to her reflective perceptions. An action is apparently irrational, furthermore, if the valued properties registered by the agent do not actually belong to the options or outcomes which she surveys or if she wrongly surveys those options or outcomes, being mistaken about their feasibility or likelihood; here the agent is in error, we may put it, about matters of value. And finally, an action is irrational, some will say, if the values registered are not suitable or objective or whatever; in this case the agent can be said to be in ignorance about matters of value.

The objection raised is fair enough. But we need not be particularly concerned, for it simply serves to remind us that practical rationality can be more or less narrowly conceived. An agent's decision-making may certainly be marred in any of the ways illustrated. And, to that extent, the agent may well be said to exhibit "practical irrationality" in her choice of action. But the form of irrationality exhibited by an agent on such occasions is not especially practical, for the failure is a purely theoretical one: it is a failure in the way she forms her judgment as to what is desirable all things considered. Our interest is in practical irrationality, more narrowly understood: if you like, in pure practical unreason. We are concerned with the failures of practical reason that can be exhibited by agents quite independently of whether their deliberations are flawed in theoretical respects.

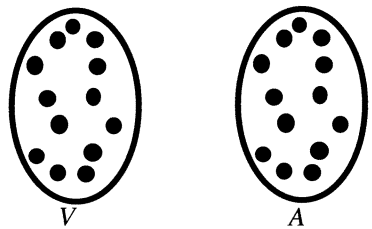
[5] The language can be misleading in other ways and needs to be used with care. See Pettit 1987.

And so we need not worry that actions and agents may not fail in this way but still count, in a wider picture, as practically irrational. We return to this topic later.

3. Practical irrationality

Given our picture of rational action and rational agency, we can now approach the question of how agents may fail in the exercise of practical reason. We approach the issue in a geometrical spirit. First we devise a geometry in which to represent rational action, then we indicate the different ways in which this geometry may be disturbed and, finally, we identify each departure from the geometry with a more or less familiar pattern of practical unreason.

First, the geometry of rational action. Imagine two closed figures or spaces, each enclosing a range of points:



Call the figure on the left the "values" space and the figure on the right the "actions" space.

Let different points in the values space represent different packages of values that might be recognised by an agent. Since we are abstracting away from the correctness of an agent's values, some points will represent what some may regard as non-values. Thus one point might represent a package comprising just the value of prudence, another just the value of beneficence, another just the value of friendship, and another just the value of fairness, while yet other points represent packages: say, packages of the values of friendship and prudence, or of the values of beneficence and prudence, or of the values of fairness and prudence. And so on. In such packages, notice, the values are unweighted.

Let different points in the actions space represent different options that an agent might choose. Think of the options as described in a way that does not reflect the values which support them; think of them as presented in a mode which makes them suitable items for the agent to control. Thus the descriptions under which a set of options present themselves remain constant, as the agent considers the different values that they promise, certainly or probabilistically, to realise. In our now well worn example, the options are to return the book or not to return it, whatever the values that the agent comes to see on either side.

In order to map rational action on a diagram constructed out of these spaces, we need to introduce two further representational devices: a broken line and a solid line. The interpretation of these devices is of the utmost importance.

Broken line.
The broken line will always connect a point in the values space to a point in the actions space. If the values point is *w* and the actions point is *b*, then the interpretation of the line is this: *given the alternative options, and given the relevant option-outcome probabilities, the value-set, w, and no set larger or smaller, leads the agent to see option b as prescriptive;*[6] *the agent weights those and other values in such a way that w supports b.* The agent may have registered many values not included in *w* but the *w*-values are those in the light of which she judges that *b* is the best thing to choose: they are the values that serve in the circumstances to make *b* seem superior to the other options. The non-*w* values which the agent may have registered will include the outweighed values present only in alternatives, and they will also include those values registered in the favoured option which did not count with the agent: the values which did not serve in the determination of the agent's judgment as to what she should do.

Solid line.
The solid line will always end at a point in the actions space and may begin at a point in the values space or at a point in between. If it connects a values point, say *x*, with an actions point, say *c*, then it means: *given the alternative options, and given the relevant option-outcome probabilities, the value-set, x, and no set larger or smaller, leads the agent to desire and choose c.* If it ends at that actions point, but does not reach back to the values space, then it means: *without regard to any valued properties—any properties represented in the V space—the agent desires and chooses c.* In order for certain *x*-values—or indeed for non-valued properties—to lead to desire and choice, the agent must have registered their presence but, as in the other case, she will have registered many other properties too. Other things being given, the generative properties are those that quicken the agent's desire for the option chosen: those that play a causal role in giving rise to a preference for that option.

[6] Two assumptions to note, both made for reasons of simplicity. We assume, first, that there is always a single option which the agent sees as prescriptive. Of course there will also be cases where the agent sees two or more options as having equal claims and, as David Lewis has reminded us, there will be cases where the agent's weighting of the values is insufficiently determinate to fix one of a number of options as the most desirable. We would have to stretch our geometrical resources in order to represent such cases, perhaps allowing a number of broken lines to originate at a given point in the values space. We assume, second, that there is no overdetermination in the relation represented by either line. We would also have to stretch our geometrical resources to represent overdetermination.

With this framework in hand, the ideal of rational action we described earlier can be represented as follows:

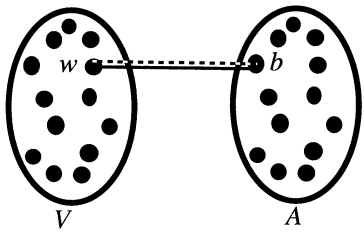


Fig 1: Reason vindicated.

In this case there is a vindication of reasons. The agent is led by certain values, *w*, to see a certain option, *b*, as prescriptive. And those same values, those same reasons, lead the agent to desire and choose *b*. Given the alternative, the fact that returning the book will fulfil my promise, leads me to see that option as prescriptive. And that very fact leads me also to desire and choose to return the book. Reason is vindicated in my action.

Even so, I may fail to be a practically rational agent; I may produce the rational action by good luck. We should remember that for an agent to be rational in making a certain choice, she must not only act rationally; she must be reliably disposed to produce such a rational action. She must not only act in a way that fits the diagram; she must be reliably disposed to act in that way. We return to this point presently.

When reason is vindicated, whether or not the agent is rational, a certain action is judged to be right in the light of certain values and then that action is desired and chosen under the influence of those values. The right action is desired, and desired for the right deliberative reasons; deliberation and desire, as we may say, resonate in harmony.

There are five ways in which this resonance may break down. The wrong action may be desired in three different ways: for the right deliberative reasons, for the wrong deliberative reasons, or for no deliberative reasons at all. And the right action may be desired in two different ways: for the wrong deliberative reasons or for no deliberative reasons at all. In the first three cases, the resonance of reason vindicated gives way to dissonance, we shall say, whereas in the other two cases it gives way to consonance.[7]

[7] Sometimes it may be to an agent's credit, by generally accepted criteria, that she displays dissonance or consonance rather than resonance: this, because the deliberative pattern with which she breaks is not particularly creditable and the achievement of narrow resonance looks unattractive from a broader perspective. Dissonance may be creditable, because it may be to an agent's credit that she is moved in desire by a property she ignores or plays down in deliberation. And consonance may be creditable for the same sort of reason. For example, it may be to an agent's credit that she is moved in desire by the consideration of just the honesty of an action when she takes account at the level of deliberation

The five ways in which the resonance of practical rationality may break down are nicely represented in the five available ways of disturbing the geometry of reason vindicated. Keep the dotted line that connects *w* and *b* in place, since all this means is that deliberation is present: there is one option that is seen by the agent, in the light of certain values, as the thing to do. There are five ways in which the solid line may then be varied, consistently with the interpretation given. Either the solid line leads to a different actions point from *b* (dissonance) or to *b* itself (consonance). If it leads to a different point, then there are three possibilities: it begins from *w* (the right deliberative reasons), it begins from another point in the values space (the wrong deliberative reasons), or it begins from somewhere in between the spaces (no deliberative reasons). If it leads to *b* itself, then there are two possibilities: it begins from a point other than *w* in the values space (the wrong deliberative reasons) or it begins from somewhere in between (no deliberative reasons).

We now go on to characterise these five forms of practical unreason. It turns out that they are illustrated by common-or-garden failures.

(i) *Reason misfires*

Agents do not always do what they take themselves to be justified, all things considered, in doing; they act in a deliberatively dissonant way. In one such case we can say that reason misfires. Certain values lead the agent to see a particular option as prescriptive but, though those same values lead her to desire and choose something, they lead her to desire and choose a different option from that which she sees as desirable all things considered. Reason misfires because, while acting on a certain set of values, she acts in a way that is not supported, in her own deliberative view of things, by those values.

This case can be represented as follows.

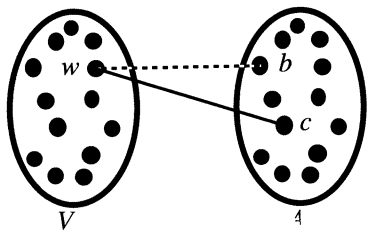


Fig 2: Reason misfires.

The value-set, *w*, leads the agent to see *b* as prescriptive, given the alternatives on offer. But that very same set of values leads her to desire and choose a different option, *c*.

---

of both its honesty and prudence: these values overdetermine her perception of the option as the right choice. Thanks to Denys Turner for a related comment.

Consider the following example, by way of illustration. Suppose I value conveying information clearly when I speak, but also value conveying that same information humorously. These values lead me to judge that, given the alternatives available, the best way for me to present a lecture is by means of a certain mix of anecdotes and formal definitions. The definitions won't do much for the humour of the occasion but they are necessary for clarity and I give clarity considerable weight, in particular more weight than humour, in my deliberations.

But now suppose that as I speak I find myself loathe to go to the definitions; I find myself sacrificing clarity to humour in a greater degree than I judge desirable. In this case, though I act on the basis of the very values that determine my all things considered judgement—clarity and humour in the conveying of information—I do not perform the action I take to be desirable all things considered. I do not convey information with the right mix of anecdote and definition. Reason misfires.

In the misfiring of reason the relative importance of the values which lead the agent to see one option as prescriptive is not reflected in the relative strength of her desires for those properties. Given the deliberative weighting of the values, one option seems prescriptive. Given the desiderative forces associated with those properties, a different option is desired and chosen. The agent's desire for clarity is too weak, or her desire for humour too strong, or a combination of these things obtains. In any case there is a failure of the balance required for rational choice.

Some have thought that an arbitrary choice is required in deciding between such descriptions of the relative strength of an agent's desires (Watson 1987). What difference is there, they ask, between the case in which the desires reflecting considerations of clarity are too weak and the case in which the desires reflecting considerations of humour are too strong? Aren't these two ways of saying the same thing? Such scepticism is misplaced.

We call a desire too strong or too weak depending on whether its strength, relative to the strength of the agent's other desires, tracks the deliberative weight of the corresponding value: its weight relative to the weight of the values to which the agent's other desires correspond. And this fact about a desire would in turn emerge in decision-making contexts, actual or counterfactual. Thus, the desire reflecting a value that is habitually defeated in action by the desires associated with other, less weighty values, is too weak; whereas the desire reflecting a value that habitually defeats other, weightier values in action is too strong.

But though reason misfires because of a failure of balance between the weights and the degrees of strength associated with certain values, this imbalance does not mean that reason will always misfire. Consider someone whose degrees of desire as between values like clarity and humour are slightly out of alignment with the relative, deliberative weights that she assigns to such properties. We can imagine that this person could make the right choice; she could choose the same option that would be chosen by the rational agent, in whom deliberative weights and desiderative forces are perfectly aligned. We return here to an observation

made in connection with the case of reason vindicated. The agent who acts in a fashion that vindicates reason may not be rational herself. She may not be reliably disposed to act in that way; she may fail to be rational by a degree which does not show up in this particular context of choice.

The diagram which represents reason vindicated marks what is in common between the rational agent and the person envisaged here: their choosing the same option. But it also enables us to bring out the difference between them. For, given that the person envisaged does not exemplify the required balance of deliberative weights and desiderative forces, there are bound to be some situations in which her reason will misfire in the manner illustrated. There are bound to be at least some counterfactual decisions where the options are such that the imbalance between the weights and forces leads her to desire and choose a different option from that which she sees as prescriptive. In those situations the choice she makes will display the geometry of reason misfiring.

(ii) *Reason internally undermined*

There is another, perhaps more familiar, case in which an agent acts in a deliberatively dissonant way, failing to do what she takes herself to be all things considered justified in doing. In this case, the agent acts, and acts on the basis of one or more of her values, but does not act on the basis of the values which lead her to see a particular option as desirable all things considered. Reason does not misfire, as the values the agent acts upon are not the very values which direct her deliberative conclusion. Reason's verdict is undermined by values ignored in the framing of that conclusion. It is internally undermined, undermined from within, in the sense that at any rate it is values, and not any more exogenous forces, which cause the problem.

We can represent this case as follows, in our second diagram. Given the alternatives, the set of values represented by w leads the agent to see the action, b, as prescriptive; but it is the set of values, x, which leads the agent to act and it leads her to choose action c, not b.

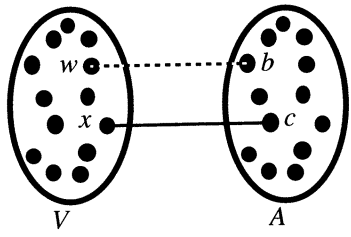


Fig 3: Reason internally undermined

Consider an example by way of illustration. Suppose I am in company and people begin to make jokes at the expense of my absent friend. Though the jokes are funny, they are also moderately hurtful, sufficiently so that a good friend would not go along with them, though not sufficiently so that a complete stranger

wouldn't rightly find them funny. In this situation considerations of loyalty support my withdrawing from the conversation altogether, letting others carry on with the jokes if they so desire, whereas considerations of humour support my going along with the joke. However, all things considered, we will suppose, loyalty presents getting up and leaving as the option to perform, the one desirable all things considered.

Imagine now that though loyalty leads me to see leaving as prescriptive, though the weight attached to that property is greater than the weight attached to the fun of going along with the jokes, the strength of my desire to be loyal is not correspondingly greater than the strength of my desire to enjoy and contribute to the humour. In this case I will stay and go along with the fun, despite my recognising that this is not the desirable option all things considered. Reason will be undermined, and undermined by considerations of the kind from which it takes its own lead. Reason will be undermined, as we say, from within. The explanation of reason's being undermined from within is that, though relative to the agent's other values, a certain value or value-set has a given weight, the desire for that valuable property does not have a corresponding degree of strength, a corresponding force, relative to the desires for the other valued properties. The possibilities divide, then, as before. The desire may be too weak relative to those other desires, or those other desires may be too strong, or the case may involve both deviations.

(iii) *Reason externally undermined*

In the cases just described, in acting contrary to what she takes herself to be justified in doing all things considered, the agent still acts upon a value she has. But sometimes agents act intentionally and knowingly contrary to what they take themselves to be justified in doing all things considered, because they act on the basis of desires that do not reflect their values at all. Sometimes agents act in a deliberatively dissonant way, without acting in the light of any properties that they value. The registering of certain properties may serve to arouse desire and choice but the properties registered do not figure as values in their deliberations; they have no weight whatsoever.

We can represent this case in our third diagram.

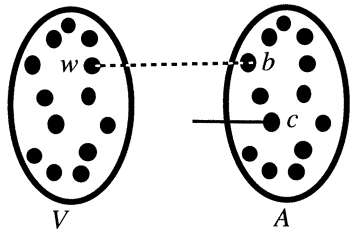


Fig 4: Reason externally undermined

The value-set represented by *w* leads the agent to see option *b* as prescriptive, given the alternatives available. But the desire on which the agent acts is not gen-

erated by those values: it is generated without regard to value, so that the solid line does not begin in the values space. And that desire leads the agent to produce action *c*, not action *b*.

We might illustrate this case by reference to the heroin addict, or indeed the distressed parent, that we mentioned earlier. The heroin addict described thinks, not just that taking heroin is undesirable all things considered, but that there is nothing at all to be said on the side of taking heroin. And equally the distressed parent thinks that drowning her baby is undesirable in every possible respect. Yet the addict takes heroin, the parent drowns her baby, and in doing these things they each act intentionally.

We describe this sort of case as one where reason is undermined from without, reason is externally undermined. Reason is undermined rather than supported, because the agent does not choose the option which she sees as prescriptive; it exhibits deliberative dissonance. Reason is undermined rather than misfiring, because the agent does not act on the values which lead her to see that option as prescriptive. Reason is externally rather than internally undermined, because the desire which produces her action is formed without regard to values of any kind. Reason is usurped by a complete outsider, a desire that reflects none of the considerations which weigh with the agent.

(iv) *Reason internally underpinned*

In most circumstances in which we act there are many reasons for doing what we judge we have most reason to do, all things considered. For many reasons converge on a single course of action in particular circumstances. This fact is, we believe, now widely accepted. It serves to explain why, for example, in adjudicating the debate between consequentialism, deontology and commonsense moral theory, we have to consider fantastic cases, not ordinary cases, in order to see how these theories differ from each other in their practical upshot.

However, even when different reasons all converge on a single course of action, there is still a question as to which reasons lead the agent to see the action as prescriptive. And, given that question, there is also the question as to whether the considerations which lead her to see the action as prescriptive are the reasons which lead her to desire and choose the action. Thus we can see room for a further failure of practical reason, albeit a relatively benign one: reason is underpinned rather than undermined, for the agent does what reason requires, even if she does it for the "wrong" reasons. Reason is internally underpinned, because it is reasons or values, and not any more foreign influences, which lead the agent to perform that action.

The case is represented by our fourth diagram. The value-set *w* leads the agent to see option *b* as prescriptive, given the alternatives available. And the agent does indeed come to desire and choose *b*. But the agent is led to desire and choose option *b*, not by the value-set, *w*, but rather by a different set, *x*.

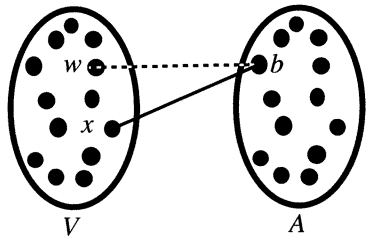


Fig 5: Reason internally underpinned

Let's consider an example. Suppose that, in certain circumstances, all things considered I have most reason to return a book I have borrowed to the person who gave it to me. Among the values that are relevant are honesty and prudence, each of which require me to return the book. Given the alternatives, the honesty considerations, and those alone, lead me to see returning the book as prescriptive. My all things considered judgement is, as we might say, "determined" by honesty, not by prudence; the prudence of returning the book plays no role in my seeing that option as most desirable.

This fact about my judgement can be captured counterfactually as follows. Imagine that I had believed that it was not prudent for me to return the book to the person who gave it to me, but still honest. Then I would still have found most reason to return the book; I would still have seen the returning of the book as prescriptive. And if I had believed that honesty had required me to give the book to someone else, though prudence still required me to return it to the person who gave it to me, then I would have found most reason to give the book to someone else; I would have seen that action as prescriptive.

But now imagine further that though I see returning the book as prescriptive in light of the honesty considerations, I do not produce the choice in light of those considerations. Rather I produce it in light of the prudence considerations, or in light of the prudence considerations combined with the honesty considerations. I do what reason prescribes but I do not do it for the reasons in virtue of which reason prescribes it. Reason is underpinned by considerations it does not invoke; it is underpinned from within.

This fact about the basis of my desire and choice can also be captured counterfactually. Suppose that it had not been honest to return the book to the person who gave it to me, though it remained prudent to do so; I might have found out she had stolen it. In that case I might still have given the book back: I would have done so if moved by prudence alone, or if moved by a combination of prudence and honesty in which prudence plays the more powerful role. Suppose on the other hand that, though honest, it had not been prudent for me to return the book. In that case I might not have given it back: I would not

have done so in the case of being moved exclusively, or in major part, by prudence.

The case we have in mind here is like the internal undermining of reason, so far as the agent is led by one set of values in her deliberation, and by another in the formation of desire and choice. The difference between the two is that the desideratively effective value-set leads in this case to the same choice, whereas it leads in the other to a different one. The action is deliberatively consonant rather than dissonant but it does not display the resonance of desire with deliberation which we associate with reason vindicated. The agent does the right thing, intuitively, but for the wrong reason.[8]

The failure involved in the internal underpinning of reason is obviously very different from either of the first three failures. It is benign rather than malign, for the consonance it secures is behaviourally indistinguishable from the case where reason is vindicated.[9] The fact that reason can be underpinned internally in this way means that there are devices imaginable whereby I can try to ensure that I behave as reason requires, or others can try to ensure this for me.

Consider once again the case where I go along with the joke against my friend. Whatever the source of the failure—whether it be a case of reason misfiring or reason undermined—it might well be that, though I am disposed to go along with the joke, I wouldn't be disposed to go along with the joke if the company knew that the person at whose expense the jokes were being told was a friend of mine. For I would then have, as I now do not, reasons of reputation to quit. They would think badly of me if I were to go along with the joke. Reasons of reputation are, perhaps, not the most admirable reasons for refraining from going along with the joke. Certainly my friend wouldn't be too pleased to find out that that is why I refrained. But these reasons might be enough to get me to do what reason requires. And so, concerned as I am with whether or not I do the right thing, I might find myself with sufficient reason to say "He's a friend of mine you know", so changing my circumstances, and thereby changing the reasons available to me for refraining from going along with the joke.[10]

[8] There are different kinds of deliberative consonance—consonance as distinct from resonance—which our approach allows us to distinguish. One sort involves the replacement of the honesty considerations, to take the example just given. In this case the honesty considerations would not have produced the behaviour on their own. They are incapable of getting me to be honest about returning the book: either they are too weak for the job, to return to a familiar dichotomy, or the considerations that argue for keeping the book are too strong. Thus it is only because the prudence considerations intervene that I am saved from an undermining of practical reason. Another sort involves the buttressing or supplementation of the honesty considerations, rather than their replacement. Here I return the book in the light of a combination of the honesty and prudence considerations. The honesty considerations would not have been sufficient on their own to get me to return the book—indeed the same may be true of the prudence considerations on their own—but the combination of both sorts of reasons is sufficient and indeed effective.

[9] Benign? Perhaps not by all lights. By some, as Mark Sainsbury has pointed out to us, "The last temptation is the greatest treason/To do the right deed for the wrong reason" (T.S.Eliot, *Murder in the Cathedral*).

[10] On this topic we are especially grateful to Jeanette Kennett for helpful conversations. For similar thoughts see Kennett forthcoming.

But not only can I try myself to ensure that I do the right thing in such cases. Others, particularly those others who have a hand in shaping the institutions of our society, may try to ensure that I, like everyone else, do what most of us recognise as the right thing. The enterprise of institutional design, an enterprise that is as old as democracy itself, is concerned precisely with ensuring that if people are not spontaneously virtuous in this or that regard, if they do not do the right thing for the right reasons, then at least they will conform to virtue's demands; they will have reasons enough of other kinds to behave as the public good requires. Such reasons may be provided, under appropriate institutional pressures, by fear of the law, fear for one's financial fortunes, fear for one's reputation, or whatever (Brennan and Pettit forthcoming).

Even republican theorists of democracy who have argued for the need for public virtue, and who have seemed to stress the need for virtue if institutional design is to be successful, have often had in mind just behavioural conformity to the demands of virtue. Thus Tocqueville writes of Montesquieu on virtue: "We must not take Montesquieu's idea in a narrow sense... When this triumph of man over temptation results from the weakness of the temptation or the consideration of personal interest, it does not constitute virtue in the eyes of the moralist, but it does enter into Montesquieu's conception, for he was speaking of the effect much more than the cause".[11]

(v) *Reason externally underpinned*

There is a second sort of case in which an agent does the right thing but not for the right reasons. This also represents a variety of deliberative consonance that falls short of resonance. In this case the agent sees a choice of action, *b*, as prescriptive in light of a value-set, *w*. But while the agent does produce behaviour *b*, she is not led to do so by the reasons provided by *w*. She produces *b* without regard to any values whatsoever. The case is represented in our last diagram.

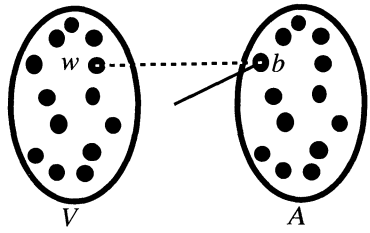


Fig 6: Reason externally underpinned.

Consider the following example. As I am walking down a footpath I see a ladder leaning up against the wall. I reflect on my reasons for walking under it as against walking around it and decide that, all things considered, I have more reason to

[11] Quoted from the preparatory notes to Volume 2 of *Democracy in America* in Aron 1968, p. 201.

walk around. Though I can't see anyone and can't see any equipment, I know that ladders are usually put up against walls like this because people are working on the roof. If there are people working on the roof then there is some chance that they have equipment near the ladder, and so there is a chance that, if I walk under the ladder, I will be hit by a falling object. Since it is only mildly inconvenient to walk around, it is best to walk around. And so I walk around. Now it seems perfectly possible that, in such circumstances, though I walk around the ladder, and so do what I have most reason to do all things considered, I may not do so for any reason that I have. I may walk around the ladder for no deliberative reason.

We have probably all been introduced to the superstition that bad things happen to people who walk under ladders. Imagine that I walk around the ladder, not because of believing the superstition—that would give me a reason, albeit a bad reason, to do so—but because my introduction to the superstition in childhood has left me compulsively and reasonlessly inclined to do so. If someone asks me why I don't walk under the ladder I have to answer: "I don't know, I just don't want to. I *really* don't want to". When I refrain from walking under a ladder because I really don't want to, in this way, then I refrain from doing so for no reason. Or so it seems to us.

Now we see how it can be that someone who does what she is most justified in doing may yet do what she does for no reason. For she may do what she is most justified in doing on the basis of a desire that in no way reflects any of her values. Reason is underpinned, as in the previous sort of case, but now it is underpinned from without; it is underpinned by a more or less brute desire. The action is deliberatively consonant but it is produced without regard to any values or reasons.

With this example described, others should readily come to mind. One obvious example is a variation on a case introduced by Donald Davidson (1980). I am in bed but remember that I forgot to brush my teeth. Suppose that given the importance of dental health to me, I see getting up and washing my teeth as the thing to do; I see the choice as prescriptive. What is perfectly possible, by analogy with the example just described, is that while I do get up and wash my teeth, I do not do so under the influence of the value of dental health. I may do so out of a compulsive feeling of guilt or discomfort at lying in bed with my teeth unwashed, a feeling laid down in the drill and training of childhood. In this example, as in the other, my reason is underpinned from without, underpinned by a force which owes nothing to the influence of values.[12]

In discussing the internal underpinning of reason we said that there are various devices whereby such underpinning is ensured and behavioural virtue is

[12] We distinguished some varieties of consonance in a footnote to our discussion of internal underpinning. In one kind the right reasons are replaced by the other reasons and in another they are supplemented by those reasons. Clearly there are corresponding possibilities here. The desire that I have without regard to value may replace the effect of the right reasons, the case being one where those reasons do not possess the force to move me to action. Or the desire may supplement the independently inadequate force which those reasons have. In order to represent this last case, we would need to extend the resources of our geometry, allowing a solid line terminating in an option to be forked, with one point of origin in the values space, the other not.

produced. A similar point obtains in this case. Much of the drill and training whereby we try to get our children to do what is right, by the values we instill in them, is likely to have the effect of inducing more or less compulsive, and perhaps guilt-driven, desires to behave in appropriate ways. A trivial example might be getting them to clean their teeth every evening, with the sort of result illustrated above. But we can easily imagine examples of a more substantive import. As institutional measures may serve to prop up reason internally, so many of the means of moral education may serve to prop it up from without.

4. Comparisons and contrasts

We have distinguished between the deliberative and the intentional perspectives on the explanation of action. We have argued that though, in the rational agent, these perspectives march steadfastly in step, in the irrational agent, they all too often come apart. And we have provided a geometry of these failures of practical reason, a geometry which directs us to the different ways in which actions can be irrational. Reason may misfire, reason may be internally or externally undermined, or reason may be internally or externally underpinned. A resonance of desire with deliberation, to invoke a different metaphor, may be replaced by a dissonance or a mere consonance.

At the beginning of this paper we said that our approach is distinguished from the established tradition of discussing practical irrationality by two features. First, it is not deferential to common sense or the philosophical tradition; it derives the different sorts of practical unreason from novel premises, rather than starting with the received categories. And second, the failures it identifies are at once distinctively rational and distinctively practical failures. We can now elaborate on these two matters. We will discuss them in reverse order, beginning with the second feature.

Some treatments of practical unreason fail to preserve anything of unreason, anything of irrationality, in the phenomena discussed. They discuss phenomena like compulsion or wantonness or weakness of will but depict them only as departures from autonomy, for example: only as failures of self-command or self-rule (Frankfurt 1988; Bigelow, Dodds and Pargetter 1988, 1990; Bigelow and Pargetter forthcoming). The idea here derives from Kant, for whom reason requires self-rule, but the interpretation envisaged loses the connection with reason. Self-rule is taken to mean just the triumph of higher-order desires: the triumph of reflexive desires as to what desires to have.

Our approach, by contrast, identifies a distinctive irrationality involved in the different ways in which deliberation and desire can come apart. If an agent judges that a certain option is to be done, if she sincerely sees that option as best, then any failure to take that judgment fully to heart is a failure of reason. It represents a failure that is continuous with the failure involved in believing certain proposi-

tions and seeing that a further proposition follows from them without being led to make any consequent adjustment in one's beliefs.

But the approach we have taken not only enables us to recognise distinctively rational failures at the origin of action; it also allow us to cast those failures as distinctively practical. Here there is a second point of contrast with the contemporary literature. In our earlier discussion we mentioned a variety of theoretical ills that might beset deliberative judgment: ignorance or error in regard to the valuable properties registered in the different options; selective or biassed attention to the valuable properties present, with the properties actually registered differing in identity or weight from those that are reflectively acknowledged; and illogic or inferential failure in the derivation of the deliberative conclusion: the conclusion derived is not actually supported, even in the agent's own lights, by the evaluative premises. Most approaches to practical irrationality assimilate failures of practical reason to one or other of these categories.[13]

Take, for example, the many different ways of understanding what is known in common sense as weakness of will. The age-old Socratic approach, under which virtue is knowledge, assimilates weakness of will to ignorance or error about relevant matters of value (McDowell 1979). And approaches that have commanded more interest in recent times assimilate it to other theoretical failures. One assimilates it to the pathology of selective or biassed attention: the agent in reflection sees the options in one way, the agent in action sees them in another, so that the deliberative judgment acted on is not the deliberative judgment reflectively endorsed (Jackson 1984, Schick 1991). And another, popularised by Donald Davidson in particular, assimilates it to inferential failure: by the agent's own lights, the evidence supports the deliberative judgment that one option is desirable—in Davidson's way of thinking, the agent judges that that option is desirable-all-things-considered, is desirable-relative-to-all-considerations—but the agent, in an inferential lapse, forms and acts on the judgment that a different option is desirable (Davidson 1980).[14]

We agree that practical reason is plagued by ignorance, error, selective and biassed attention, and inferential failure. Not only that. We also think that these sorts of failure are of great importance and that the literature which characterises them makes an enormous contribution to our understanding of practical unrea-

[13] An exception is Michael Stocker, whose approach we find congenial. See Stocker 1979.

[14] Susan Hurley (1989, Chs. 7 and 8) introduces an interesting variation. Under this approach, as under ours, the inferential input to deliberative judgment is a registering of valued properties, such that any one property can present an option as desirable *pro tanto*, and can continue to present it as desirable *pro tanto*, even after the option is seen as not desirable *simpliciter*. Here there is a contrast with Davidson (1980), for whom a prima facie consideration in support of an option—the counterpart of the *pro tanto* support—ceases to provide any support if, all things considered, the option does not appear to be desirable. Given the apparatus of *pro tanto* reasons, one which we essentially endorse, Hurley argues that weakness of will is characterised by acting on a *pro tanto* judgment of desirability rather than a judgment of desirability *simpliciter*.

son; what reservations we have bear on matters of detail.[15] But we believe that, however serious, the maladies characterised in this literature do not exhaust the ways in which practical reason may break down and, more particularly, that they neglect the breakdowns of pure practical reason: the breakdowns which are not distinctively theoretical in character. All the maladies discussed affect the final, deliberative judgment of desirability: the judgment of desirability, as we think of it, all things considered. But the breakdowns of pure practical reason, the breakdowns which are not particularly theoretical, bear on the connection between the final judgment of desirability and the agent's desire, not on the status of the judgment of desirability itself. And it is such breakdowns that are identified and taxonomised in the approach we have taken here.

We all have a powerful, pretheoretical intuition that human agents can be fully cognisant of, and fully sensitive to, the reasons which support their performing one action, and yet go on and perform another. The distinctive feature of our approach is that it supports this intuition. We recognise all the failures to which the ordinary approaches draw attention but we give countenance to other failures of reason as well: failures of pure practical reason, failures which occur without any lack of cognisance or sensitivity on the part of the agents.

So much for the substantive contrast between the approach taken here and more standard approaches to practical unreason. The other feature which marks off our approach is methodological in character rather than substantive. It consists in the fact that we are not deferential to the categories of common sense, or of the philosophical tradition, in delineating the possibilities of practical unreason; we derive a taxonomy of failures from novel rather than received premises.

The literature on practical unreason emphasises a variety of departures from practical reason or, understood in a narrow sense, virtue. An agent can depart from virtue by displaying mere continence, for example, or by being weak of will, or compulsive or capricious. The main divide is between departures from virtue that result in a right action from the deliberative point of view, as with continence, and departures that result in a wrong action, as with weakness of will, or compulsion or caprice. Can our schema substantiate these distinctions? We believe that it can.

Weakness of will, compulsion and caprice are all instantiated both in cases where the agent does the wrong thing for the wrong deliberative reason and in cases where she does the wrong thing for no reason at all. So what is the difference between them? With all three phenomena, there is a mismatch between the degrees of desire present in the agent and the values that she recognises in delib-

[15] One matter is worthy of particular notice here. Even under inferential failure, the rational response will be to desire and choose the option that is seen as prescriptive. The rational response will have to match the mistake made in the deliberation with a mistake, if you want to call it that, in the generation of desire and choice. But the rational agent's dispositions are unlikely to be able to produce the desire and choice required by an inferential mistake in a reliable way. And so the agent may find herself, happily, incapable of living up to her judgment. This sort of inability should be educative, as remarked by Alison MacIntyre (1990) and Jeanette Kennett (forthcoming).

eration. The difference between the phenomena, so we conjecture, relates to the causation and character of this mismatch.

A difference of causation marks off caprice from the other two pathologies. With weakness of will and compulsion, the mismatch is something visited upon the agent from without: it is a legacy of her nature, her past or whatever. With caprice, that is not so: the mismatch is something for which she, as she is at the moment, is blameworthy; it involves a more or less wilful departure from reason. As a difference of causation marks off caprice, so a difference in the character of the mismatch marks the divide between weakness and compulsion. Roughly, we think that it is appropriate to ascribe weakness of will when the mismatch is one that the agent is capable of handling: recognising where her desires are leading, she is capable of inhibiting their effect, say by reflecting on the long-term, more or less egoistic costs of following them. We think that it is appropriate to ascribe compulsion rather than weakness of will to the extent that this contemporary sort of self-control is not possible: to the extent that the agent is enslaved by the desires that move her away from the path prescribed in deliberation.

So much for actions that are compulsive or weak or capricious. Finally, we trun to continence, in particular the continent agent. This type of agent is traditionally taken to be someone who struggles to do the right thing, and generally succeeds, making distinctive efforts of self-mastery or self-management: efforts which the virtuous agent does not need to make. Does our schema allow us to make sense of this picture? We believe it does.

Continent actions must come out, on our approach, as right actions done for the wrong reasons. It is natural to assume, then, that the continent agent is someone who produces such continent actions and produces them non-accidentally or reliably. And that assumption explains why the continent agent fits the traditional image. If an agent is reliably to produce the right action, but not for the right reasons, then she must rely on providing herself with special incentives, or more or less blind habits, to get her moving in the right direction. She must equip herself with resources which will wring from her a compliance that is not in her nature. She must make up for a lack of spontaneous virtue by becoming a successful tactician in the art of self-management. In a word, she must conform to the received image of continence.[16]

We hope that these remarks are enough to show that although our taxonomy of pure practical failures of reason is generated by a non-trivial distinction between deliberative and intentional dimensions, although it does not start out from the received wisdom on practical unreason, it does serve to make sense of received categories. The taxonomy directs us to different failures from those that are generally emphasised in the current literature, and connects equally well with the long, partly common-sensical tradition of thinking about practical irrationalities.

[16] Thanks here to Mark Johnston for a helpful comment.

5. Conclusion

We mentioned in passing that Kant introduced the idea that non-heteronomy is required for practical rationality. This idea gave rise to the characterisation of all forms of practical unreason as varieties of government from without, government by something other than the self. One way of summing up the approach adopted here is to show how it gives new, non-Kantian life to this political imagery.

Our concern has been with narrow practical rationality, as we have stressed. So what should such rationality involve, in terms of the metaphor of non-heteronomy? What should narrow non-heteronomy be taken to require? The received line would say, autonomy: the rule of the self—the rule of the *autos*—rather than an alien rule. But on our approach the natural response is to say that in the narrow sphere of practical rationality, non-heteronomy is not *self*-rule or autonomy; it is *right* rule or "orthonomy" (Pettit and Smith 1990, p. 588).

What is wrong with heteronomy, on our approach, is not that it involves the rule of the "heteros" in the sense of the exogenous; what is wrong with it is that it involves the rule of the "heteros" in the sense of the inappropriate. We see the non-heteronomous agent, the agent who is practically rational in the narrow sense, as someone in whom desire is appropriately governed, not just as someone in whom the government of desire is exercised by her. Thus we take a very different view of non-heteronomy from post-Kantian existentialists like Sartre who require any operative desires to be affirmed in an act of radical choice (Sartre 1957, Part 4, Ch. 1). And equally we see things very differently from someone like Harry Frankfurt, who requires any operative desires, or at least any operative ground-level desires, to be desires that are endorsed a level up: desires that the agent desires to act on (Frankfurt 1988). Our image of non-heteronomy is driven by a more traditional metaphor of good government than the democratic metaphor which seems to inspire such visions. The good government of desire is a regime under which desire is faithful to the rule of deliberation; being endogenously inspired and maintained is not enough, even if it is necessary.

The notion of orthonomy, however it contrasts with post-Kantian ideals, connects up with the tradition which emphasises the requirement of executive virtues in a rational agent. The non-executive or substantive virtues require an agent to be a lover of the good; the executive virtues require her to be a good lover. Examples of virtues that are predominantly, if not exclusively, executive include temperance, courage, fortitude, and an impartiality across times and persons: if you like, justice. As we see such virtues, they are requirements or aspects of orthonomy. To be orthonomous requires a temperance about the things that can let loose uncontrollable desires; a courage which does not let the desire for one's own welfare excessively warp one's choices; a fortitude which enables one to bear up under adversity, maintaining a desiderative connection with the things one values; and an impartiality which keeps the claims of one's future self, and the claims of other persons, as powerful in the generation of desire as more immediate counterparts.

Such executive virtues are derived in the Aristotelian tradition from a preference for the middle way. It is not surprising that our notion of orthonomy should be seen as a generalised version of the executive virtues for, intuitively, the ideal of orthonomy represents a version of the Aristotelian doctrine of the mean. Or at least it does to the extent that the doctrine bears on the narrow matter of how an agent desires rather than the broader question of what she desires.[17] The important thing is not to assume control of one's desires, as in the existentialist or quasi-existentialist vision. The important thing is to be someone in whom desires are neither too strong nor too weak. It is to be someone in whom desires are generated by values, and generated with forces equivalent to the weights that those values are accorded in deliberation. The forces must not fall short of the weights, nor must they rise in excess of them. The forces and the weights must be in balance.

Our conception of narrow practical rationality thus gives us at least one reason, if we are to stick with the Kantian imagery, for taking non-heteronomy as orthonomy rather than autonomy. But there is also a more general consideration which favours this rendering. The ideal of right government may be understood narrowly or broadly, depending on how far we are prepared to specify the goals of the governors. As we have characterised orthonomy, it describes only a narrow ideal of practical rationality: an ideal of pure practical reason. But that narrow ideal fits naturally into a broader one: an ideal under which desire answers to deliberation, as the narrow ideal requires, and deliberation itself escapes theoretical defects like illogic, inattention, error and ignorance; an ideal under which the agent is substantively as well as executively virtuous. The fact that orthonomy can be understood narrowly or broadly means that as an ideal of pure practical reason, it is continuous with a fuller and more rounded picture of practical rationality. As an ideal of pure practical reason, it proclaims its incompleteness on its face; it does not suggest, as the ideal of autonomy has sometimes done, that it represents the be-all and the end-all of morality. And that, surely, is to its credit.[18]

Australian National University,
Canberra, ACT 2601,
Australia.

PHILIP PETTIT

Monash University,
Clayton, Victoria 3168,
Australia.

MICHAEL SMITH

[17] See Urmson 1980, especially the summary on p.163. But see also Hursthouse 1980-81.

[18] We are grateful to Geoffrey Brennan, Richard Holton, Lloyd Humberstone, Jeanette Kennett, Rae Langton, Peter Menzies and Mark Sainsbury for helpful comments. We are also very grateful for the many helpful comments that we received at presentations of the paper, especially at its initial presentation in a seminar on "Value in Action", held at Monash University in August 1991.

REFERENCES

Aron, Raymond 1968: *Main Currents in Sociological Thought*, Vol 1. Harmondsworth: Penguin Books.

Bacharach, Michael and Hurley, Susan, eds., 1991: *Essays on the Foundations of Decision Theory*. Oxford: Oxford University Press.

Bigelow, John, Dodds, Susan and Pargetter, Robert 1988: "Against the Will". *Pacific Philosophical Quarterly*, 69, pp. 307-24.

——1990: "Temptation and the Will". *American Philosophical Quarterly*, 27, pp. 39-49.

Bigelow, John and Pargetter, Robert forthcoming: "Autonomy and Integrity". Monash University, mimeo.

Brennan, Geoffrey and Philip Pettit forthcoming: "Hands Invisible and Intangible". *Synthese*.

Charles, David and Lennon, Kathleen, eds., 1992: *Reduction, Explanation and Realism*. Oxford: Oxford University Press.

Dancy, Jonathan, ed., forthcoming: *Reading Parfit*. Oxford: Basil Blackwell.

Davidson, Donald 1980: "How is Weakness of Will Possible" in his *Essays on Actions and Events*. Oxford: Oxford University Press.

Flanagan, O. and Rorty, A.O., eds, 1990. *Identity, Character and Morality*. Cambridge, Mass.: MIT Press.

Frankfurt, Harry 1988: "Freedom of the will and the concept of a person", in his *The importance of What We Care About*. Cambridge: Cambridge University Press, 1988.

Hurley, Susan 1989: *Natural Reasons*. New York: Oxford University Press.

Hursthouse, Rosalind, 1980-81: "A False Doctrine of the Mean". *Proceedings of the Aristotelian Society*, 81.

Jackson, Frank1984: "Weakness of Will". *Mind*, 93, pp.1-18.

——1985: "Internal Conflicts in Desires and Morals". *American Philosophical Quarterly*, 22, pp.105-14.

Johnston, Mark 1989: "Dispositional Theories of Value". *Proceedings of the Aristotelian Society*, Supp.Vol. 63, pp. 139-74.

Kennett, Jeanette forthcoming: "Mixed Motives". *Australasian Journal of Philosophy*.

MacIntyre, Alison 1990: "Is Akratic Action Always Irrational?", in Flanagan. O. and Rorty, A.O., eds., 1990.

McDowell, John 1979: "Virtue and Reason". *Monist*, 62, pp. 331-350.

Pettit, Philip 1987: "Utilitarianism without Universalisability". *Mind*, 96, pp. 74-82.

——1991a: "Decision Theory and Folk Psychology", in Bacharach and Hurley, eds, 1991, pp. 147-75.

——1991b: "Realism and Response-dependence". *Mind*, 100, pp. 587-626.

Pettit, Philip and Michael Smith 1990: "Backgrounding Desire". *The Philosophical Review*, 99, pp. 565-92.

——forthcoming: "Parfit's P", in Dancy forthcoming.

Rorty, A.O., ed., 1980: *Essays on Aristotle's Ethics*. Berkeley: University of California Press.

Sartre, Jean Paul 1958: *Being and Nothingness* (tr. H.Barnes). London: Methuen.

Schick, Frederic 1991: *Understanding Action*. Cambridge: Cambridge University Press.

Smith, Michael 1987: "The Humean Theory of Motivation". *Mind*, 96, pp. 36-61.

——1992: "Valuing: Desiring or Believing?", in Charles and Lennon, eds., 1992.

Stocker, Michael 1979: "Desiring the Bad: An Essay in Moral Psychology". *Journal of Philosophy*, 76, pp. 738-753.

Urmson, J.O. 1980: "Aristotle's Doctrine of the Mean", in A.O. Rorty, ed., 1980.

Watson, Gary 1977: "Scepticism about Weakness of Will". *Philosophical Review*, 86, pp. 316-39.

——1982: "Free Agency", in his *Free Will*. Oxford: Oxford University Press.