JULIA TANNEY

DE-INDIVIDUALIZING NORMS OF RATIONALITY

(Received 11 January 1994)

INTRODUCTION

It seems to be a platitude that what makes behaviour irrational is its failure to accord with some particular norm of rationality and it seems right to say that intentional action by and large conforms to these norms. These considerations might encourage one to attempt to explain an individual's ability to act rationally, and account for some of her lapses, by attributing to her "knowledge" – either explicit or tacit – of what the norms require. The norms of rationality in some sense govern thought and action. But is the sense in which they do this captured by construing them as psychologically internalized rules, or as causal determinants of behaviour?

The need to attribute some particular principle of rationality to an individual is defended by Davidson explicitly in his characterization of akrasia.[1] I should like to explore his attempt to "individualize" the principle, or render it into a norm which is cognized by the individual whose actions are governed by it. This will require taking some space to explicate Davidson's causal account of intentional action, which, of the sake of making the arguments clear, I'll just accept. I'll show that it is not necessary to individualize a principle of rationality in order to characterize an individual's actions as internally irrational. In the second half of the paper I'll develop this argument by considering in detail what explanatory role an individual's cognitive grasp of such norms might play. I'll argue that there is no construal of "cognitive grasp" such that attributing cognitivist grasp of a norm to an individual would explain her dispositions to act in accordance with what the norm prescribes, either directly, or via her second-order explicational abilities. I argue in

Philosophical Studies **79:** 237–258, 1995.

the end that cognizing a norm of rationality could only be considered constitutive of an individual's ability to *obey* it on a very artificial and stipulative sense of "obey". I conclude that it is a mistake to construe the principles of rationality as norms or rules which may or may not be obeyed or followed.[2]

I. MENTAL STRUCTURE

In "Actions, Reasons, and Causes" (*ARC*) (reprinted in Davidson, 1980), Davidson argues that rationalizing explanations, which explain an agent's actions by citing his reasons for doing what he does, are a species of causal explanations. We can specify the mental cause of the action – the reason for which an agent acts as he does – by citing a pro-attitude or desire the agent has towards actions of this kind (or toward the events or states brought about by actions of this kind), and a belief that the action is of this kind or will bring about the desired outcome.

In *ARC* Davidson suggests that the logical relationship between the contents of reasons and actions can be described as a practical syllogism: the pro-attitude and belief that cause the action (my desire for adventure, and my belief that spending the weekend in Barcelona will enable me to satisfy this desire) have contents that can be described as premises in an argument the conclusion of which is a description of an intention that corresponds to, or prescribes, an action. Thus the desire-constituent of a reason (say, my reason for spending the weekend in Barcelona) can be expressed as: "Any act of mine which is likely to yield adventure is desirable". When coupled with the belief premise "Spending the weekend in Barcelona is likely to yield adventure", it follows that any act of mine which is my spending the weekend in Barcelona is one I may judge to be desirable. The problem with this view, is that any act-type which is likely to yield adventure will be one whose desirability follows from my pro-attitude. So, in "How is Weakness of the Will Possible?" (*WoW*) (reprinted in Davidson, 1980), Davidson proposes that the propositional content of the desire-component should be changed to express what he calls a "prima facie" evaluation (or judgment) which qualifies event- or action-types: any action of mine is desirable insofar as it is likely to yield

adventure. This, together with a premise expressing the belief component: "Spending the weekend in Barcelona is likely to yield adventure", implies a judgment which must itself be relativized as the major premise is above: "Spending the weekend in Barcelona is desirable insofar as it is likely to yield adventure". But this kind of judgment is too weak to determine action: although spending the weekend in Barcelona is desirable insofar as it is likely to yield adventure, it might be undesirable for other reasons. In *WoW* Davidson claims that the action itself must correspond to something stronger than a "prima facie" evaluation that the act is desirable in a certain respect; it must correspond to an unconditional or all-out, singular judgment expressing the desirability of a particular action.[3]

The upshot of this view is that there will be a logical gap, according to Davidson, between the content of prima facie evaluations and the content of all-out judgments. The contents of my reasons for acting deductively imply the content of a relativized judgment on my part, that insofar as the act would enable me to satisfy a particular desire, performing it is desirable. But they don't deductively imply an "all-out", "derelativized" judgment that I ought to perform the action (say, that I ought to spend the weekend in Barcelona). And yet, this sort of all-out judgment just is the intention to act that Davidson holds to be a necessary concomitant to action.

Prima facie evaluations that conflict, like desires that conflict, must compete in order to be realized in action, so the psychological model needs to be extended to capture the logical structure between judgments made in deliberation: the relationship between judgments that adjudicate between conflicting prima facie evaluations. This structure is evinced in deliberation insofar as choices are made about which reasons are to subserve others. If these choices are rational then according to standard theories of decision, and arguably commonsense, deliberation proceeds by one's ranking or weighting certain desires or judgments higher than others on the basis of one's (rational) grasp of their relative importance and the probability of the possible outcomes.

Although deliberation (the mental process) perhaps only occurs self-consciously, a self-conscious weighting of alternatives need not be required for intentional action, nor for actions to be rationalized. All

that is required for this is that the explanatory relations between reasons, judgments about the relative ranking of reasons, and intentions that seems to be exhibited in self-conscious deliberation be attributable to the subject as part of our rationalizing project. Thus we may say that, although not self-consciously deliberating, the agent has nonetheless acted intentionally. And this seems undeniable. There is manifest rationality in all sorts of "automatic" human behaviour which is not a product of self-conscious deliberation.[4]

In *WoW* Davidson suggests that the judgment that manifests the relative ranking of reasons – i.e., the result of deliberation – is an "all things considered judgment". An all things considered judgment is "doubly relativized" (Pears, 1984). First it is relativized according to the way in which the desire would be satisfied in the commission of the action (say, as in the prima facie judgment: "Spending the weekend in Barcelona is desirable insofar as it is likely to yield adventure"). It is relativized also according to its place with respect to other desires and in light of the agent's beliefs, principles, and values. This judgment might be something like: "In light of my ranking the opportunity for adventure over prudential concerns, and in light of my beliefs about what spending the weekend in Barcelona will involve, and so on, spending the weekend in Barcelona is desirable". According to Davidson, this all things considered judgment is conditional in form and thus, like the singly relativized judgments that logically precede it, does not entail the kind of judgment which is a necessary concomitant to intentional action. Again, this latter judgment – which Davidson identifies as an intention – must be unconditional, or derelativized. So the logical gap that exists between the contents of prima facie evaluations, or sentences describing them, and the contents of intentions, or sentences describing them, is still preserved on the extended model between all things considered judgments and actions. The move from a doubly relativized judgment like "Assuming that I have considered all relevant things, I ought to spend the weekend in Barcelona" to an unconditional (derelativized) judgment like "I ought to spend the weekend in Barcelona" is not a move that is prescribed by first-order logic since, presumably, some piece of relevant information not considered might always defeat the claim that I ought to spend the weekend in Barcelona. Thus, the failure

to make such a move in one's thinking cannot (yet) be taken to exhibit a kind of logical inconsistency.

It is precisely this kind of failure that Davidson takes to be exhibited in paradigmatic akratic action: "the action of an agent who, having weighed up the reasons on both sides, and having judged that the preponderance of reasons is on one side, then acts against this judgment" (Davidson, 1982). Since this type of irrationality is, for Davidson, *ineliminable*, the introduction of the all things considered judgment is a propitious addition to the account of intentional action since it allows the logical space for akrasia. But leaving the logical space for irrationality makes characterizing the error that is manifested in it that much more difficult. This is apparent to Davidson who wonders, "since logic cannot tell me which to do, it is unclear in what respect either action would be irrational".[5]

This sets up the major programme of "Paradoxes of Irrationality", which is to diagnose which principle has been violated in akrasia, and – the thesis I'll be challenging – to characterize the norm violation intra-individually or intra-psychically so that it is reflected as an inconsistency within (descriptions of) the agent's mental states or events (including relevant descriptions of action). It is not obvious straight off how this strategy is to succeed: if practical reasoning isn't straightforward explicable in terms of logical coherence among contentful states, then it isn't obvious that surd practical reasoning can be straightforwardly explicated in terms of logical incoherence among contentful states. But worse: if such a general account were to be given, then since the right kind of logical relations are, according to Davidson, in some sense constitutive of the very thoughts and actions we're using in our descriptions of irrational states or events, there is a constant worry that the kind of inconsistency needed to characterize irrationality will at the same time undercut the identification of the mental events that are used in its description. This, then, is the paradox of irrationality. Anyone attempting to give a general account of irrationality has to show how descriptions of irrational events are possible given the apparent constitutive nature of the principles that are violated.

The burden of the discussion that follows is to explore the way in which these principles are constitutive.

II. PARADIGMATIC AKRASIA AND THE PRINCIPLE OF CONTINENCE

Before offering an account of what goes wrong in akratic action, Davidson gives an example of straightforward intentional action, in which a person acts for reasons:

> A man walking in a park stumbles on a branch in the path. Thinking the branch may endanger others, he picks it up and throws it in a hedge beside the path. On his way home it occurs to him that the branch may be projecting from the hedge and so still be a threat to unwary walkers. He gets off the tram he is on, returns to the park, and restores the branch to its original position. Here everything the agent does (except stumble) is done for a reason, a reason in the light of which the corresponding action was reasonable, for example, given that he wants to remove the branch and he believes that getting off the tram will enable him to remove the branch, it was reasonable for him to get off the tram.

Now consider additional information which renders the action akratic. When the man returns to the park to remove the branch, although he has a reason for getting off the tram, say he has a better reason (not wasting the time) to continue on it. Suppose the man accepts that he has a better reason to stay on the tram, and that he judges, all things considered, that he ought to continue. What needs to be explained, according to Davidson, is not why he gets off the tram (we saw that he has a reason for that) but rather why he doesn't act according to his better judgment.

Davidson's suggestion as to how to characterize the kind of irrationality that occurs in the paradigmatic case of akrasia is to posit a principle of practical reasoning that will bridge the logical gap between the penultimate outcome of deliberation – namely, an all things considered judgment – and the intention itself. The principle says that I ought to act in accordance with my all things considered judgment (and form derelativized judgments or intentions consistently with it). Davidson calls this a *second-order* principle, presumably because it speaks about the deliberation process itself, and doesn't necessarily get bandied around within it. Introducing the second-order principle allows Davidson to diagnose what goes wrong in a case of akrasia by pinpointing the norm that has been flouted.

I would like to take a closer look at Davidson's introduction of the second-order principle. For the sake of argument I'll accept it as diagnostic of one kind of irrationality. I would like to explore what an

individual agent's relationship to it is supposed to be, and what work this relationship is doing in the characterization of akrasia.

What follows is Davidson's formal description of what he takes to be a paradigmatic case of akratic action.

> Pure internal inconsistency enters only if I also hold – as in fact I do – that I ought to act on my own best judgment, what I judge best or obligatory, everything considered ... A purely formal description of what is irrational in an akratic act is, then, that the agent goes against his own second-order principle that he ought to act on what he holds to be best, everything considered. It is only when we can describe his action in just this way that there is a puzzle about explaining it. If the agent does not have the principle that he ought to act on what he holds best, everything considered, then though his action may be irrational from our point of view, it need not be irrational from his point of view – at least not in a way that poses a problem for explanation. For to explain his behaviour we need only say that his desire to do what he held to be best, all things considered, was not as strong as his desire to do something else. (Davidson, 1982)[6]

What does it mean to "hold" the principle? In what sense does the agent *have* the principle that he act on what he holds best, everything considered? When Davidson says that in acting akratically the agent goes against his own second-order principle, he seems to be suggesting that the principle be attributed as the content of one of the subject's own mental states.

Now it might be thought that in insisting that the paradox of irrationality only arises for individuals who "hold" the principle their actions violate, Davidson is attempting to call our attention to a distinction invoked in moral psychology between *external* and *internal* irrationality. A person is externally irrational if she violates the norms of her community in choosing some end, but internally rational if she engages in appropriate means-ends reasoning to achieve it. Thus a person might be externally irrational insofar as she want to die, but internally rational insofar as her suicide plans suit the intended goal. The idea seems to be that unless the principle of rationality is attributed to an individual as one of her ends, her violating it cannot be construed as problematic from her own point of view. But this reasoning is odd, since attributing the principle as an end presupposes the individual to whom it is attributed is already disposed to act in accordance with it, and if she is, it isn't clear what attributing it to her as an end would explain.[7]

And yet, clearly the principles of rationality may be possible objects of thought: I'm presumably thinking about them now as I write. But what of the akrates? Perhaps it makes sense to attribute to him knowledge of the principle if he is able to explicate his actions in light of it: to diagnose his akratic actions as irrational, to correct his actions which violate it, or to justify those which accord with it. And perhaps it might be thought that having these second-order abilities is necessary for him to be internally irrational. But, intuitively, this doesn't seem to be true.

To bring this point home, and to see how external irrationality with respect to the principle of continence tends to collapse into internal irrationality, suppose the agent we're considering hasn't got the ability to correct his actions in light of perceived violations of the norm, he hasn't got the ability to justify his actions in light of it, etc. Then the purported explanation of the action of this individual – which Davidson claims would prevent its characterization as (internally) "irrational" – is that "his desire to do what he held best all things considered, wasn't as strong as his desire to return to the park". But how would appealing to the strengths of the desires circumvent a diagnosis of internal irrationality? On a simple model of deliberation, whatever difference to the outcome of deliberation the strength of a desire makes, it makes before the all things considered judgment is formulated. If the desire to return to the park is stronger, it ought to be the survivor of deliberation and championed by the all things considered judgment. If instead the desire to stay on the tram survives deliberation, then presumably this is because it was stronger. But all this is true of any deliberator: even one who isn't able to correct himself when he fails to act in accordance with his all things considered judgment. Thus it isn't clear how adverting to the strength of the desire in the case where the principle can't be attributed as a representation allows us to cancel a diagnosis of irrationality; the fact of a particular desire's strength must already have been taken into account by the time the all things considered judgment was made. In fact, this is the same point that Davidson makes when he describes the error in deliberation for one who does hold the principle of continence ("the desire to return to the park entered twice over"), only here, supposing we can imagine an individual who doesn't have the explicational abilities,

we can pinpoint the same kind of failure; thus whether or not he holds the principle of continence is immaterial. Again: that a desire overpower an all things considered judgment that issues from a contest that the desire itself was too weak to win, evinces some kind of error. That it does manifest an error is what Davidson himself points to in describing the fault that occurs in his paradigmatic akratic act. And yet its being an error doesn't depend on the agent's being able to correct himself in light of failing to act as his deliberations dictate, etc. The implication that we can accept *as not (internally) irrational* the action of one who doesn't have these abilities is wrong.[8]

The upshot is that attributing the principle to an agent as an object of knowledge or as a cognitive state isn't necessary to diagnose internal irrationality. Because supposing someone does *not* have the second-order explicational abilities (that attributing knowledge of the principle ostensibly would explain), how could we *ever* escape the conclusion that nonetheless, given his status as an agent, a deliberator, a practical reasoner, etc., he *ought* – **by his own lights** – to correct himself in light of his understanding that he has acted against his better judgment? After all, what is the point of his deliberating if he isn't going to act in accordance with his deliberations? Indeed, why would he get as far along in the deliberation process as to reach the all things considered judgment if he will not act in accordance with it?

The argument is that on the hypothesis that an individual hasn't got the second-order abilities to correct and justify her actions in light of the principle, she is nevertheless acting *internally* irrationally when her actions fail to conform to it.

I'd like to backtrack now, and explore the premises in the argument above in more detail. I began this paper by stating two intuitions. First, that it seems to be a platitude that irrationality is the violation of some norm of rationality. And second, that the principles of rationality in some sense govern thought and action.

Perhaps we need the principles as diagnostic tools, but if the above argument about the principle of continence is correct (and generalizable to other principles of rationality) then it would seem unnecessary to attribute them as objects of cognition to the agent's we're attempting to diagnose or teach.

The argument rests on the premise that in order to be attributed as objects of cognition, the norms must be serving some explanatory role in the cognitive/psychological life of the individual to whom cognitive grasp of the norm is attributed. I'll argue in detail in what follows that there is no *normative* role for the cognized principle to play.

III. EXPLAINING NORM-CONFORMING BEHAVIOUR

A. *Explicit Representation*

It might be thought that attributing knowledge of a norm to an individual might explain the individual's disposition to act in accordance with it. After all, as Davidson supposes, the akrates must be disposed to act in accordance with this norm, even though on occasion he violates it.

But the fact that ones acts in conformity with a norm or principle doesn't yet give a reason to attribute knowledge of it to the agent. Your resting in your chair instead of floating away manifests behaviour that accords with the principles of physics, but attributing to you knowledge of these principles doesn't explain why you don't float away. Of course, it might explain other things, for example, it might explain how it is you're able to answer a question correctly on a physics exam about why don't float away.

Although this is a point which arises over and over again[9] it will be worth making the problem explicit in its present incarnation.

Crispin Wright (1989) describes the intuitive idea of how norms or rules might figure in a psychological explanation:

> Correctly applying a rule to a new case will, it is natural to think, typically involve a double success: it is necessary both to apprehend relevant features of the presented situation and to know what, in the light of those apprehended features, will fit or fail to fit the rule.

But he has left something out. Correctly applying a rule to a new case will involve a third success: the ability to implement the (cognized) rule in action. Thus, a cognitive explanation of a subject's ability to act rationally would require attributing to the subject grasp of a rule which she implements in action based on her recognition of what the

rule requires in a particular situation. It will obviously involve other, related kinds of successes as well – more on this in a moment.

But if my implementation of the principle of continence, say, is needed to move me from all things considered judgment to action, then why isn't a higher-order principle of continence needed to tell me how I am to implement the principle of continence non-akratically, and so on?

To see more specifically how this problem arises, consider the proposal that we attribute the principle as the content of a judgment. Now consider what the status of such a judgment should be; what role would it play in mediating between other mental states and action? We can try to answer this by borrowing from the Davidsonian model sketched in the first section. Is my holding the principle of continence, for example, tantamount to my having a pro-attitude toward my acting in accordance with my all things considered judgment? This is suggested when Davidson calls the principle of continence a second-order principle: like other principles (and values) it might express a prima facie pro-attitude although unlike other principles, this one is second-order because it would express a pro-attitude about how the deliberation process itself ought to proceed (viz., that the first-order all things considered judgment ought to issue in consistent first-order intentions).[10] If it is to be construed in this way, then we can reasonably ask what is the relation between this pro-attitude and any subsequent implementation of the principle. It cannot be a determinate relation prescribed by logic, since the most I can conclude from the principle-qua-pro-attitude is: "Any action which is the formation of a first-order intention is desirable insofar as it accords with my all things considered judgment". But we're still left with the same logical gap between this (second-order) judgment and the kind of all-out judgment needed to motivate the second-order action (in this case, the formation of an all-out first-order intention that the principle of continence was invoked to provide). And in order to fill in this gap we would need something like a third-order principle of continence, and so forth.

Perhaps the principle of continence is the content of an all things considered judgment. Then I judge, all things considered, that I ought to act in accordance with my all things considered judgments. Now the internal regress is explicit. If holding the principle (judging that I

ought to act in accordance with my all things considered judgment) were explanatory of my rational abilities at all, it would only be if the connection between my all things considered judgments and my actions were presupposed. But this was precisely the connection that the principle was invoked to explain.[11]

It ought to be clear that this is going to be a problem with respect to other abilities besides the ability to implement the result of decision-making in action. Other possible fault lines through the practical reasoning process or the logical structure of deliberation which involve perceptual, conceptual, and judgmental abilities will be affected. For each of the possible fault lines we'll presumably need a principle to govern the appropriate non-irrational passage; but each of these principles will in their turn be presupposed in the norm-obeying process or in the logical structure of a norm-obeying explanation. Presupposed, then, in the very type of explanation on offer will be our abilities to make appropriate judgments regarding the role of the norm and its applicability, its scope, its relevance, and it overridingness. This is because these judgments and abilities, in turn, will presumably be subject to errors of their own. That this is so is evident when we realize that irrationality (self-deception) affects thoughts, and judgment-formation as well as action, just as wishful thinking affects perception. The principle of continence, then, will have sister principles governing intentional abilities to cover these other possible fault lines in between, and including, perception and action. But all of these abilities are presupposed in the logical structure of rule-following explanations. From grasping the rule, and circumstances to which it applies, through weighting conflicting principles, interests, plans, and desires, down to intending to implement the rule in action and finally acting in accordance with this intention and implementing the rule in action. All of these abilities are rational abilities, they themselves admit of errors and are thus themselves subject to norms which govern their use. But not just some norms or other. Precisely the same norms that are being considered. Thus none of principles that govern abilities which are presupposed in norm-obeying can be attributed to individuals as objects of thought in order to provide a psychological (norm-obeying) explanation of the disposition to act in accordance with them.[12]

Earlier I mentioned that attributing to me knowledge of the laws of physics might explain my ability to answer correctly a question about why I don't float away – even if it can't explain why I don't float away. Perhaps the analogy can be extended to the norms of rationality. Would attributing to me knowledge of a norm explain my second-order ability to justify my actions that accord with it and diagnose and correct my actions that do not?

But my ability to justify my actions is an ability which itself is, unfortunately, subject to precisely the same threat of irrationality: it is within my justifications that self-deception is evinced. But if these second-order abilities consist in, among other things, a disposition to act in accordance with rational norms in implementing the second-order rules governing justification, then attributing to me knowledge of the norms couldn't explain my ability to justify, correct, and guide my actions in light of them for the same reason as before: because the type of explanation proposed presupposes the disposition which is part of that which is to be explained.

B. *Tacit Representation and Implicit Realization*

Our paradigm examples of norm-obeying behaviour are when the individual is able to refer to a norm – perhaps she names or describes it – in guiding her deliberations, in justifying her actions, or in diagnosing what went wrong. Perhaps it is this (partly linguistic) ability that would tempt us to attribute to her the norm as an explicit representation. But, as I've urged, *correcting*, *guiding*, and *justifying* are all themselves rational abilities and presuppose the very same dispositions, and so cannot explain them. This is all very clear in the so-called "self-conscious" cases.

But what if the norm is a *tacit* representation? Might it figure as part of a cognitive explanation nonetheless? Even if the agent cannot self-consciously pick out that stretch of reasoning that the norm governs and self-consciously guide, justify, or correct the transition in light of her conception of the norm, we might, someone might argue, have evidence that she has got the requisite second-order explicational abilities even if she isn't self-conscious of them. And her tacit knowledge of the norms

might explain these (sub-cognitive) abilities without threat of regress. Whether or not obtaining the requisite evidence is possible is a question I'll put to one side. Nevertheless, supposing it is possible, unless we had evidence that the agent *conceptualizes* the norm and evidence that it is in virtue of this conceptualization that she acts in accordance with it, we wouldn't be able to distinguish between her merely acting in accordance with the norm and her having the abilities which would be required for (presumptively) attributing tacit knowledge of the norm to her.

But even if obtaining the requisite evidence for the second-order abilities were possible, it should be evident that the same regress problems arise. Whether or not the subject can refer to the norm in her self-conscious explications, if we are to be granted license to attribute (even tacit) knowledge of it to her, her behaviour manifesting these second-order sub-cognitive abilities needs to be sufficiently robust. These abilities will involve "sub-perceiving" the norm, "sub-perceiving" the stretch of behaviour to which it applies, "sub-judging" its applicability, and "sub-acting" in light of it. Nevertheless, to tacitly "sub-perceive", "sub-judge", and "sub-implement" it *correctly* as opposed to *irrationally*, presuppose the very same dispositions, operating sub-cognitively, that these tacit representations are, on the hypothesis we're considering, posited to explain.

There is another ostensible way of effecting the appropriate internal connection which might be thought to avoid the regress difficulties. That is, if the second-order justificatory or explicatory abilities which involve the individual's ability to discriminate the norm and guide action in light of it are somehow "implicitly realized" or "causally determined" in the individual to whom "knowledge" of the norm is attributed.[13]

But now we face a dilemma: either what is implicitly realized is the subject's *cognitive grasp of a norm* or it isn't. That is, either the subject manifests abilities which are complex enough presumptively to warrant attributing cognition of the norm to her (or one of her subsystems); in which case this case devolves into the case of tacit representation of the principle. And, again, a tacit representation of the principle cannot explain norm-conforming behaviour since the disposition to behave in conformity with the norms is presupposed in attributing to the individual the tacit representation, since she or her subsystems have to perceive

and implement the norm *correctly*; that is, in conformity with the very same norm of rationality.

Or the subject doesn't manifest such complex abilities; in which case the case devolves into one in which she merely acts in accordance with the norms. But we would have no more reason to attribute the norm to her as part of her cognitive "hardware" to explain the fact that she acts in accordance with it than we have reason to attribute laws of physics to her cognitive "hardware" to explain the fact she acts in accordance with them. We might describe her as acting in a way that is subsumed by laws – but in doing so we're making no hypothesis about her psychological or cognitive processes. And this gets at the intuitive difference between norms and laws. Norms, unlike laws, are *prescriptive* and it is this fact that tempts one to attribute them to the individual as cognitive states or representations in the first place, since doing so at least seems to explain the role they play in *guiding* the individual to whom they are attributed. But attributing them as causal determinants of behaviour preempts this function. And this brings us to a correlative problem of conceiving the norms as instantiated in, or as causal determinants of, behaviour. Just as their role as prescriptions for guiding behaviour would be lost, so, too, would their role in diagnosing rational error, or motivated irrationality. But the norm was introduced to characterize how the individual who has violated it has erred.

It seems that the principle of continence cannot discharge two functions simultaneously: it cannot both be used as diagnostic of irrationality and as that which is causally instantiated in, or determines, behaviour. Notice that here we run up against the paradox of irrationality as described earlier – and this is a point which prescinds from specific issues about cognition: if the right kinds of relations between mental properties or states are constitutive in this strong sense – if it is necessary that intentional phenomena are either logically or causally determined by norms of rationality that govern them – then irrationality cannot be diagnosed by adverting to the lack of such a relation. In acting akratically, for example, a person has nonetheless acted for reasons and in that sense has acted intentionally. Thus, the sense in which the akratic action is irrational cannot be described as its failure to issue via a relation which is at the same time held to be *necessary* for intentional action.

The point obtains for any proposed principle of rationality: if acting in accordance with it is considered necessary for an action's being intentional, or if its implementation is necessary for a thought's having content, then failing to act in accordance with it, or its failure to be implemented, will not be a possible characterization of an irrational action or thought.[14]

These arguments have important ramifications. There is a strong intuition that we need to make out an *internal* connection between norms and the individual who acts in accordance with them in order to make sense of the intuition that she acts *because* of the norms. A disposition to act in accordance with the norms doesn't seem to give us the right kind of non-contingent relation required for explanation. But, I argue, this relation cannot be made out as a *cognitive* one such that the norms themselves are objects of knowledge or desired ends and a person engages in reasoning to implement or satisfy them. This is because the "reasoning" here will presuppose the disposition that attributing these very norms was meant to explain. It would seem, the, that if my disposition to act in conformity with the norms of rationality is, indeed, some kind of achievement, it isn't a *cognitive* achievement.

IV. CONSTITUTING NORM-OBEYING ABILITY

What of the intuitively plausible idea, then, that the norms of rationality *may be* possible objects of thought? Perhaps attributing to me knowledge of a norm of rationality doesn't *explain* my rational abilities either directly, or via second-order explicational abilities, by the arguments above; but perhaps my having knowledge of the norms *consists* in my ability to justify my actions. And perhaps my having this second-order ability is necessary for me to be considered truly rational. If so, maybe we can make out the sought after "internal" connection after all. My *following* a rule or *obeying* a norm, as opposed to my merely acting in accordance with it, might *consist* in my ability to justify my actions in light of the principle prescribing it. And perhaps I can only be considered to be a truly rational agent if I am able to *follow* the rules of rationality

in this sense. Correspondingly, I am only akratic when I have the ability to *correct* my actions in light of the perceived violation of the norm.

But what could it mean to *justify* an action in light of a norm of rationality? We can make sense, in general, of justifying an action by citing a rule: I might, for example, justify a certain move in chess by citing the rule which allows it. But it would be odd to say that I can justify a rational action (against the incursion of irrationality) by citing a norm of rationality, since the abilities that I'm using in the second-order action are precisely the ones with respect to which my first-order action is allegedly being justified. This isn't so for the chess case. Someone might ask me to defend my move, and subsequently might ask me to defend my interpretation of the rule that I cite in justification of it, but in so defending my interpretation, *I'm not making a move in chess*. But in the case of rationality, I'd be making the same kind of move both times. And if my first-order rational dispositions need to be defensible by me in order for me to be considered truly rational, why aren't my second-order dispositions in need of the same justification? They invoke, again, the very same moves that allegedly need to be justified in the first place.

Note, finally, a move which is *not* open to those insisting that the ability to *follow* a rule of rationality is necessary for true agency. They are not allowed to appeal to the way our practice does function to argue that – whether or not justification runs out eventually – citing a principle of rationality is what we do, *in fact*, count as justification.[15] Because we don't in fact advert to these principles nor require that anybody do so. Although we learn to cite justifications for all kinds of actions, in so doing our rationality is assumed.

CONCLUSION

What general conclusions can be gleaned from the arguments above? They suggest that the strategy of construing norms of rationality as cognitive objects, rules to be followed, or norms to obeyed, is misguided. I argued in the first part of the paper, using Davidson's principle of continence as an example, that individualizing the norm is not necessary for characterizing a subject's thoughts or actions as irrational by her

own lights: whether or not attributing the norm to an individual is explanatory or constitutive of her explicational abilities, what would matter for characterizing her as irrational in light of her own standards is that she act against the norm. But if she has all of the abilities required for us to diagnose her as acting against it in the first place, the fact that she acts against it is sufficient for her to be violating a norm which is in some sense her own. If the principles of rationality were to play a normative role for the individual then her inability to express awareness of these norms in explicating her behaviour would itself be a sort of failure. But not one which would neutralize an akratic action.

Do they play such a normative role? In the second part of the paper I argued that construing the principle of continence as represented in a psychological or cognitive state of an individual took us nowhere in explaining her disposition to act rationally: either directly or via possible second-order explicational abilities. Internal regresses thwarted attempts to conceive of the principle of continence as explicitly represented, and as tacitly represented. I argued also that this result will threaten norms of a similar ilk: namely those governing perceptual, conceptual, and judgmental abilities that are presupposed in cognitive, rule-following explanations. And although the regress difficulties might be obviated by attempts to construe the "instantiation" of the principle as a causally necessary condition for intentional action, doing so renders the norms explanatorily inert. If we need them at all, we need them as diagnostic of irrationality. But if they're to be construed as causally instantiated in behaviour, they've metamorphized into laws and their functions as norms have disappeared.

Indeed, it is the attempt to conceive of the instantiation of an norm of rationality as either a causally or logically necessary condition for intentional action that will render irrational action paradoxical since if the instantiation of a particular norm is a necessary condition for intentional action or for thought, then irrationality cannot be described as the failure of the instantiation of that norm.[16]

Finally I considered the suggestion that attributing knowledge of the principle to an individual is necessary insofar as it constitutes the individual's ability to justify his rational action. I rejected this idea on the grounds that an individual doesn't have the ability to justify his

rational actions, in any non-stipulative and non-artificial sense of the term.

But in what sense, then, do the norms of rationality govern thought and action if they are not properly construed as objects of cognition? The answer is that they set up the practice of ascribing thoughts and action. This is a point often made by Davidson in discussions about the principle of charity. The principles of rationality seem to play the same kind of role. They are not rules or norms that figure in our attributive practices. They are presupposed by it. But if they ground the practice of interpretation, it would be a category mistake to explain features of the practice by individualizing them.[17]

NOTES

[1] A correlative move – the need to internalize reasons – is made in Davidson's causal account of intentional action. In Tanney (1995), I consider how the arguments I give here affect the thesis that reasons are causes.

[2] I've attempted to draw out, in more detail, how Davidson's strategy is subject to the problems suggested in Baier, 1985. In so doing I hope to be meeting many of Pears' (1985) criticisms of Baier's line. A detailed exploration of these suggestions is important, since many of the ideas defended by Davidson – notably, the idea that reasons are causes – (which, agreeing with Baier, I think is the real culprit in rendering irrationality a paradox) have been dominant (again) in the philosophy of mind for the last 30 years, largely because of Davidson's arguments. The intuition that there is something odd about this thesis, and perhaps something correspondingly odd about the research programmes that depend on it, needs to be developed.

[3] That an all-out judgment about the desirability of the action is necessarily connected to an act performed intentionally is implicit in *WoW*. The connection between all-out judgment and intending is made in "Intending" (reprinted in Davidson, 1980). For criticism of the former, see Pears, 1984.

[4] This move plays a crucial role in the argument and in similar arguments of the kind. It is to insist on the explanatory propriety of attributing intentional properties in a way which may not correspond to the subject's (self-) explicational abilities as manifested either in her avowals or in her acts of self-interpretation. Burge makes an analogous move in his anti-individualism arguments when he insists on the explanatory propriety of attributing concepts to an individual whose concept-explicational abilities are incomplete. Note that Davidson sanctions this move in Davidson, 1985a.

[5] This is the other crucial move of the argument and arguments of its kind: it is to insist on the explanatory propriety of attributing intentional properties in a way which

manifests a kind of internal error. In this case actions are identified as akratic or irrational, in such a way that they are nevertheless tokens of intentional action. Burge makes an analogous move in his anti-individualism arguments when he argues for the propriety of attributing concepts to individuals whose application of them is partially mistaken.

[6] At times, it seems that Davidson *stipulates* that the "paradigmatic" akratic action that he wants to analyze necessitates attributing the norm violated to the akrates, when for example, he says: "The standard case of akrasia is one in which the agent knows what he is doing, and why, and knows that it is not for the best, and knows why. He acknowledges his own irrationality". At other times as in the quote cited in the text above, it seems he is *arguing* for the necessity of the attribution. In any case, I hope to return a negative verdict to a possibility he contemplates at the end of this article: "I have urged that a certain scheme of analysis applies to important cases of irrationality. Possibly some version of this scheme will be found in every case of 'internal' inconsistency or irrationality" (p. 305, 1982).

[7] I argue for this in detail below. In later articles (1985b and 1985c), Davidson suggests that no one can be interpretable as failing to hold the principle but he seems to see this as no threat to his argument for requiring that the principle be attributed to the individual in order to diagnose akrasia. If no one can be interpretable as failing to hold the principle,then, as I'll argue below, the norm has metamorphized into a *law*, and there would be no reason anymore to attribute the principle to any particular individual (a fortiori as part of any particular individual's "cognitive equipment").

[8] A number of people have suggested that on a more complex model of deliberation, this "fault" can be modelled so as to allow the possibility of desires surviving in action, despite deliberation to the contrary. But this misses the point. The question is not whether the phenomenon (in which the "strongest" desires survive in action despite deliberation to the contrary) is *possible*; the question is whether or not the phenomenon is *irrational*. My claim is simply that if it is irrational, we don't need to attribute cognition of the norm that is violated to the subject in order to see it as an internal error.

[9] See, for example, Carroll (1895), Wittgenstein (1953), and Sellars (1963).

[10] Davidson's comments about the agent not having a *reason* for acting against his all things considered judgment also sustain the interpretation of the principle of continence as a second-order prima facie principle. His comment here implies that such a reason is possible. So, a second-order deliberation process might be set up for deciding whether or not to act on the principle of continence in the first-order case. This would give the (second-order) principle the status of a prima facie judgment.

[11] Suppose, on the other hand, that one tries construing the principle as represented in a second-order standing intention expressing the pro-attitude that I ought to act in accordance with my first-order all things considered judgment, whose adoption necessitates implementation of the principle in action. This move corresponds with conceiving the principle as implemented in causal/psychological processes, which will be discussed in more detail shortly.

[12] It should be clear that rule-following explanations are not being rejected, *tout court*.

Attributing rules that govern a practice as representations may be explanatory of a person's ability to participate in it as long as certain abilities (namely the ones that are presupposed by rule-following explanations) are not themselves part of the phenomena to be explained by them. Thus my following the rules of chess might be explanatory of abilities I have to make allowable moves in the game, but in this case, the perceptual, conceptual and intentional abilities which constitute the ability to follow a rule are not the abilities to be explained. If this line of reasoning is correct then anyone interested in giving a cognitive/psychological rule-following type explanation must presuppose that we are rational agents in order to attribute principles which would explain a person's actions; it must be presupposed that we act in accordance with our deliberations (in this case that we not only grasp what the rule requires of us but also implement this understanding in action).

[13] Although many have argued that it isn't really knowledge in this case, since it can't be considered an achievement, nor for that matter, can it explain a cognitive ability, the cognitive sciences are full of pleas to allow usage of "knowledge" to slacken and to allow the notion of "rule-following" to slacken too, on the grounds that cognitivists are using it with such success. But it is precisely my aim to begin to question whether their use of it is necessary.

[14] It is this argument that threatens Davidson's causal account of international action. It also suggests a problem with any account that attempts to realize normative relations in causally necessary conditions for action, thought, or meaning.

[15] This move is arguably available to those who want my following *modus ponens*, say, on a logic exam to constitute my ability to get the answers right.

[16] The paradox might be interpreted as the following argument in which seemingly acceptable premises lead to an unacceptable conclusion: 1) The instantiation of rational norms is necessary for intentional actions. 2) But irrational actions are nevertheless intentional actions. 3) Irrationality is the violation of a rational principle. 4) Therefore, irrational actions are both possible and impossible. A paradox is generated if the first premise reads that for every norm, its instantiation is necessary for intentional action, which is the reading, I suggest, that an individualistic account of rationality forces on us given positive explanatory constraints on mental attribution.

[17] This paper has benefited from many comments. I'd especially like to thank David Velleman, Crispin Wright, Barry Loewer, and Frank Döring for comments early on, and Hartry Field, Christopher Peacocke, and Jim Hopkins for recent discussions.

REFERENCES

Baier, 1985. "Rhyme and Reason: Reflections on Davidson's Version of Having Reasons", in Lepore and McLaughlin.

Davidson, 1980. *Essays on Actions and Events*, Clarendon Press, Oxford.

Davidson, 1982. "Paradoxes of Irrationality", in R. Wollheim and J. Hopkins, eds., *Philosophical Essays on Freud*, Cambridge University Press, Cambridge.
Davidson, 1985a. "Reply to Bratman", in B. Vermazen and M. Hintikka, eds., *Essays on Davidson Actions and Events*, Clarendon Press, Oxford.
Davidson, 1985b. "Deception and Division", in Lepore and McLaughlin.
Davidson, 1985c. "Incoherence and Irrationality", *Dialectica*, Vol. 39, No. 4, pp. 345–354.
Carroll, Lewis, 1895. "What the Tortoise Said to Achilles", *Mind*, Vol. IV, pp. 278–280.
Lepore, E. and B. McLaughlin (eds.), 1985. *Actions and Events – Perspectives on the Philosophy of Donald Davidson*, Blackwell, Oxford.
Pears, David, 1984. *Motivated Irrationality*, Clarendon Press, Oxford.
Pears, David, 1985. "Rhyme and Reason – Response to Baier", in Lepore and McLaughlin.
Sellars, 1963. "Some Reflections on Language Games", in *Science, Perception, and Reality*, Routledge & Kegan Paul, London.
Tanney, J, 1995. "Why Reasons May Not Be Causes" *Mind and Language*, Vol. 10, Nos. 1–2, pp. 103–126.
Wittgenstein, 1953. *Philosophical Investigations*, trans. by G. E. M. Anscombe, Basil Blackwell.
Wright, Crispin, 1989. "Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics", in A. George, ed., *Reflections on Chomsky*, Oxford, Basil Blackwell.

Department of Philosophy
University of Kent at Canterbury
Canterbury CT2 7NY
UK