Weakness of Will and Divisions of the Mind

Edmund Henden

Some authors have argued that, in order to give an account of weakness of the will, we must assume that the mind is divisible into parts. This claim is often referred to as *the partitioning claim*. There appear to be two main arguments for this claim. While the first is conceptual and claims that the notion of divisibility is entailed by the notion of non-rational mental causation (which is held to be a necessary condition of weakness of the will), the second is explanatory and claims that the notion of divisibility is required for the causal explanation of weak-willed action. In this paper I want to argue that the partitioning claim remains unsupported, no matter how it is interpreted, and that weakness of the will can be made perfectly good sense of without the idea that the mind is divisible into parts. In fact, there are available various explanatory models each of which characterizes different psychological mechanisms that may be involved in weakness of will, none of which depends on any claims about mental division. I describe three familiar mechanisms and argue that weakness of will may occur as the result of any one of them.

The structure of this paper is as follows. I begin in section 1 by distinguishing three possible interpretations of the partitioning claim. In section 2, I present a conceptual argument for mental partitioning and argue that it fails. In section 3, I examine the claim that there are explanatory reasons for the partitioning claim. There are at least two ways of supporting this claim, depending on how the partitioning claim is interpreted. I argue that neither is successful. So where does this leave us as far as the explantion of weakness is concerned? In section 4, I argue that there is little point in trying to identify one unique causal mechanism that explains all cases of weakness of the will. Instead, we should be open to the possibility that a variety of causal mechanisms may play a role; alone, or in combination with others.

1

The question of interpreting the partitioning claim is essentially an ontological question about the identity conditions of 'mental parts'. What is a 'mental part', and how can it be distinguished from other 'mental parts'? There are various possibilities, depending on the criteria of individuation chosen. Mental parts may be individuated in terms of *structure* (i.e. what sort of mental entities they are composed of and the interrelations between these entities), in terms of *causal role* (i.e. what mental and physical changes they bring about), or possibly in terms of

European Journal of Philosophy 12:2 ISSN 0966-8373 pp. 199–213

some combination of these criteria (say, a structure of a certain type that brings about certain mental or physical changes). Closely connected with choice of criteria will be a certain view of the temporal permanence of mental parts, i.e. the time period they occupy in a persons mental life (whether they go out of existence almost as soon as they come into existence, or whether they have a more permanent presence). Since Davidson's seminal paper 'Paradoxes of Irrationality', there have been suggested at least three different ways of individuating mental parts (including Davidson's own).

(a) Mental parts as time-dependent control structures

This interpretation of the partitioning claim goes back to Davidson's original proposal in 'Paradoxes of Irrationality'.[1] Although the notion of 'mental part' appears to play a prominent role in Davidson's argument, he does not say much by way of characterizing what he means by this term. The following is based on what I take to be the most natural interpretation of his remarks.

Let us start with structure. Do mental parts have a particular structure that distinguishes them from other mental parts? According to the (a)-interpretation, a mental part is a structure of beliefs and desires that is internally rational, and that is an independent source for the derivation of a syllogism in practical reasoning. In other words, it is *a reason*. However, it is not the fact that it has this particular structure that individuates it as a *separate mental part*. Clearly, the mind consists of infinitely many beliefs and desires which, viewed from a certain perspective, are internally rational, but which do not qualify as separate mental parts. Rather, what individuates mental parts on the (a)-view, is their role in causing irrationality. In Davidson's words: 'The breakdown of reasons-relations defines the boundary of a subdivision'.[2] I take this to mean that some local structure of interrelated beliefs and desires can be viewed as a semi-independent mental part only if it causes a mental state or an action it does not rationally justify. For example, if an incontinent desire to perform some action causes the agent to perform that action even though she judges that she has better reasons for not performing it, that desire and its supporting beliefs (and perhaps related desires) can be viewed as a semi-independent mental part.

Given this view of mental parts, the structure centred around the incontinent desire must do various things in order for the irrational causation to be successfull. First, it must make the agent ignore principles of rationality, such as *the principle of continence* which says that one should always do what one judges is best all-things considered. If the agent does not ignore this principle, she will perform the continent action. Second, it must actively prevent the bulk of the agent's reasons from producing the continent action since it seems perfectly conceivable that these reasons could retain enough 'momentum' to produce the continent intention even if the principles of rationality was left inoperative.

Now this view of the causal role of mental parts has consequences for their temporality. It suggests that the mental part centred around the incontinent

desire does not need to have any permanence beyond the time it takes for the agent to perform the incontinent action. Another feature of mental parts, on the (a)-view, therefore, appears to be that mental parts may come into existence for a brief moment, take control over the agent, and then fade out as soon as she has acted.

(b) Mental parts as sustained goal-structures

This interpretation of the partitioning claim has been proposed by Sebastian Gardner.[3] According to Gardner, it is not breakdown of reason relations that individuate mental parts. Rather, what is essential are the various features of the mental sub-structures; the relations which exist between these structures and between the entities which belong to them. Thus, according to Gardner, mental parts exist independently of irrational mental causation. Instead, they are sustained features of the mind. They are organized around more or less permanent goals of the agent, each of which may have supporting structures of beliefs, desires, expectations, assumptions and attitudes. What individuates these parts on this view is the relation of *cohesion* between members of each structure, in Gardner's words, 'the centripetal force exerted by the internal agreement of a set of propositional attitudes', combined with the relation of *mental distance*, or non-integration between members of different structures.[4] The concepts of cohesion and distance are conceptually connected according to Gardner, in the following way: 'Internal cohesion necessarily contributes to mental distance: if mental item A coheres with B but not C, and B is inconsistent or conflicts with C, then there will be a tendency for A not to integrate with C. And mental distance in turn implies the existence of internally cohesive structures, for holistic reasons: if an isolated mental item fails to integrate with the bulk of the mind, this must be because it derives support from other mental items, whose magnetism helps to draw it apart'.[5] I shall return to an example of how this may work in a case of weakness in section 3, but let me now move on to the last interpretation of the partitioning claim I shall consider.

(c) Mental parts as rational centres of agency

A version of this interpretation of the partitioning claim is defended by David Pears.[6] The (c)-interpretation differs from both the (a) and the (b) interpretations in claiming that the semi-independent mental parts are not just structures of propositional attitudes centred around some desire of the agent, but *rational centres of agency*. What does this mean? According to the (c)-theorists it means that the interactions between the different mental parts should be understood in terms of the strategic interactions between rational individuals. The idea is that mental parts, just like rational individuals, have a capacity for planning, i.e. consider various strategies to achive their goal in the most effective way. For example, a structure centred around an incontinent desire may use manipulative

strategies to distract the rational main structure from focusing on reasons against its favoured practical conclusion, i.e. distort these reasons or even fabricate 'new' reasons in support of the favoured conclusion. In this way the interactions between the various sub-structures of the mind can be viewed on the model of the game-theoretic interactions between rational agents.

How are the mental parts individuated on this view? Just as the (a)-theorists, the (c)-theorists argue it is the breakdown of reasons-relations that defines the boundary of mental parts. But in contrast with the (a)-theorists, they do not take such breakdowns to consist in a mental state irrationally causing another mental state or action. The problem with the latter view, they claim, is that it does not entail that there has to be a conflict in the agent's mind. But if there is no conflict, if the agent does not *believe* that the mental causation violates a rational constraint, what need is there of a semi-independent mental part?

Instead of interpreting the breakdown of reasons-relations in terms of irrational mental causation, the (c)-theorists see it as produced by the failure of a mental state to interact in a rational way with some element belonging to the rational main-structure. For example, the agent's *cautionary belief* that the causation of a certain mental state or action would violate a rational constraint would, under normal conditions, be sufficient to prevent this state or action from occurring.[7] But sometimes this belief may fail to prevent the irrational state or action from occurring. As a result of this failure, a split is created in the agent's mind. In other words, in contrast with the (a)-theorists, the (c)-theorists allow for the possibility of cases of irrational mental causation which do not require mental partitioning, i.e. cases in which one mental state irrationally causes another, but where there is no failure involved, that is, where the agent does not believe that she is being irrational. Since the cautionary belief (in principle) may be prevented from intervening in the main-system for any duration of time, the sub-structure that contains it must have a sustained existence beyond the performance of the incontinent act. In this respect, the (c)-interpretation bears a certain similarity to the (b)-interpretation, which also emphasizes the sustained temporal existence of the mental part.

I shall now examine some of the arguments which have been used to justify the partitioning claim.

2

The weak-willed agent freely, deliberately and intentionally performs a particular action A against her judgement that some incompatible action B, would be better. Consider now the following simple argument for the partitioning claim:

(1) If weakness of the will is possible some reasons are not the reasons for what they cause.
(2) If some reasons are not the reasons for what they cause, the mind must be divisible.

(C) If weakness of the will is possible, the mind must be divisible.

A version of this argument can, I believe, be found in Davidson.[8] Let us just call it 'the conceptual argument' for the partitioning claim. Let me first say a few words about the premises of this argument.

In premise (1) it is claimed that 'some reasons are not the reasons for what they cause'. What does this mean? The most reasonable interpretation, I think, is that some reasons may *cause* the agent to draw an incontinent conclusion associated with a weak-willed action, but fail to *rationally justify* this conclusion in the light of the biggest set of relevant reasons she has considered. However, even if they fail to rationally justify the conclusion, they are still *reasons* for it, in the sense that they give it *some* support. It just so happens that they have been outweighed by other of the agent's reasons.

If this is a correct reading of premise (1), it seems to me to be true for the simple reason that there does not appear to be any other way of explaining how the weak agent can self-consciously draw an incontinent practical conclusion. On the one hand, her drawing of this conclusion cannot be rationally explained by the biggest set of relevant reasons she has considered, since if it was, her conclusion would be rationally justified by those reasons, in which case it would not be incontinent. This suggests the presence of some causal influence other than her reasons. On the other hand, since she acts freely and intentionally, the cause that explains her drawing of the incontinent conclusion cannot be completely external to her reasoning (like a sudden impulse). If it was, the weak-willed action associated with this conclusion would not have been a full-blown intentional action.[9] In fact, the cause must *itself* be *a reason*; it must be *her reason* for drawing the conclusion, although, once again, it fails to rationally justify this conclusion. But how is that possible? The only way seems to be if the causal strength of a reason may be out of line with its rational weight in the agent's practical reasoning, the consequence being that it causes her to draw the incontinent conclusion, even if it should not from the point of view of the biggest set of relevant reasons she has considered. Hence premise (1).

Let me move on to premise (2). Note first that there are two possible readings of this premise. Either it can be taken as the strong claim that non-rational mental causation *in general* implies a divided mind, or it can be taken as the weaker claim that it *only* implies a divided mind in cases of weakness of the will. While the latter allows for the possibility that non-rational mental causation can occur without a divided mind, the former requires that a divided mind is postulated in every case of non-rational mental causation.

Few would, I think, seriously argue for the general version of this claim.[10] For it seems clear that non-rational mental causation is something that can occur in a variety of circumstances, many of which do not require a divided mind. Davidson mentions an example of someone who tries to remember a name by humming a certain tune.[11] Here there is a mental cause that is not a reason for that which it causes, but it is not needed to postulate a divided mind to account for such cases. The difference between this case and cases of weakness is that in

the latter kind of case the mental effect is *irrationally produced* by the non-rational mental cause. Even though humming a tune in order to recollect a name is a mental cause but no reason for recollecting the name, it is not *irrationally* producing the recollection of the name. We should, therefore, take premise (2) as a claim about non-rational mental causation *specifically* as it occurs in cases of weakness of the will.

Now suppose we take (2) in this sense, why should we believe that the mind must be divisible? One reason, due to Davidson, is that the notion of irrational mental causation leads to a dilemma unless we accept that the mind is divisible.[12] Let us call this 'the dilemma argument' for mental partitioning. What is the dilemma? The idea appears to be the following: thought of as *causes*, the mental causes of weak-willed actions belong in the category of the non-rational, or to borrow a phrase of John McDowell's, 'in the logical space of nature'. To place something in the logical space of nature is to situate it in the realm of law and the natural sciences. As such mental causes seem unsuitable for explaining weakness since they threaten to reduce it to *non-rationality* rather than irrationality. On the other hand, if these same mental causes are thought of as *reasons*, they belong in the category of the rational, or to borrow McDowell's phrase, in 'the logical space of reasons' which is constituted by normative relations such as one thing's being warranted by another.[13] But this also seems to make them unsuitable for explaining weakness since now they threaten to reduce weakness to *rationality* rather than irrationality.[14] How can this dilemma be avoided? Davidson claims the only way is to assume that the mind is divisible into parts. He explains why in the following passage:

> There is, however, a way one mental event can cause another mental event without being a reason for it, and where there is no puzzle and not necessarily any irrationality. This can happen when cause and effect occur in different minds. For example, wishing to have you enter my garden, I grow a beautiful flower there. You crave a look at my flower and enter my garden. My desire caused your craving and action, but my desire was not a reason for your craving, nor a reason on which you acted. (Perhaps you did not even know about my wish.) Mental phenomena may cause other mental phenomena without being reasons for them, then, and still keep their character as mental, provided cause and effect are adequately segregated.[15]

Davidson's suggestion is that once we apply the idea of segregation to a single mind, the dilemma can be avoided.

Now does the conceptual argument give us a good reason to accept the partitioning claim? The main problem, I think, is premise (2), which is supposed to be supported by the dilemma argument. There are two parts to this argument. First, there is the claim that the notion of irrational mental causation gives rise to a dilemma. Second, there is the claim that this dilemma only can be avoided if we assume that the mind is divisible. Let us assume for the moment that the first part

of this argument is correct; in other words that it is correct that the notion of irrational mental causation gives rise to a dilemma. Why does this show that *the partitioning claim* is true? At this point the dilemma argument appeals to a supposed analogy between the interaction between different minds and the interactions between different parts within one single mind; since the notion of non-rational mental causation can be understood in the former case, the idea is that we can use it as a model for understanding irrational mental causation in the latter case as well. The trouble is that there is another and simpler way to understand this analogy which does not assume systemic separation in cases of weakness. All we need assume is *distance* in the causal chain between mental cause and effect. Consider once again Davidson's example: in the flower case, my desire to have you enter my garden is not your reason for entering my garden although it is the cause of your entering my garden.[16] Why? Because it causes your entering *indirectly* through causing your noticing of the flower in my garden. The analogy to a case of weakness is then as follows: my incontinent desire to do A against my judgement that some incompatible action B would be better, is not a reason for doing A although it is the cause of my doing A. Why? Because it causes my doing A *indirectly* through causing my ignoring of various rational constraints (such as the principle of continence). In other words, we can speak in both cases of *a reason* that is not a reason for that which it causes, but that is because it causes its effect *indirectly* rather than directly. So far no reason has been given for saying that mental cause and effect must occur in *different parts of the mind*.

But what about the first part of the dilemma argument, the claim that the notion of irrational mental causation gives rise to a dilemma of the type described? Even if the analogy can be interpreted in a way that does not support the partitioning claim, it could still be true that the notion of irrational mental causation leads to a dilemma which *somehow* could be remedied if we accept mental partitioning?

If there is a dilemma here, it is only on the surface and it certainly does not require measures as drastic as introducing separate parts of the mind to avoid it. The appearance of a dilemma trades on an ambiguity in the notion of 'reasons' between normative (good) reasons on the one hand, and motivating (possibly bad) reasons, on the other. While a normative (good) reason may rationally justify a certain action or attitude from the agent's own point of view, it may still not be this reason that motivated the agent to act. The reason that motivated her to act may, from the agent's own point of view, fail to provide rational justification, not in the sense that it is not *a reason* for the relevant action *at all*, but in the sense that it has been outweighed by her reasons for not performing it.

Now the claim that there is a dilemma is based on the idea that a mental cause, thought of as *a reason*, cannot explain weakness because it belongs in the category of the rational, and what counts in 'the logical space of reasons' is justification, or being justified. This is true if by 'reason' we mean *a normative* reason that provides justification of the action from the agent's own point of view. Of course, taken in this sense, an irrational mental cause cannot explain weakness since if it

did, it would rationally justify its effect which would be inconsistent with, at the same time, irrationally producing this effect. But if by 'reason' we mean *motivating* reason, possibly a very *bad* reason, that is, a reason that has been outweighed by other of the agent's reasons, the conclusion that the mental cause thought of as *a reason* cannot explain weakness, does not follow. Of course, taken as a *motivating* reason, a mental cause can be a reason for an action even if it produces it irrationally; it is a reason for it in so far as it gives it support and motivates it. There is no dilemma here. Of course, the fact that the agent's reason in a case of weakness is a motivating reason does not fully explain why she performed the action. We also need an explanation of why it caused her to perform this action, that is, why it was *causally stronger* than its rational weight would suggest. But that is a question about explanatory mechanisms (to which I shall return in the conclusion), and it is not clear why the partitioning claim would be relevant here.[17]

To sum up. The conceptual argument for the partitioning claim seems to rest on an unsupported premise, namely the idea that irrational mental causation gives rise to a dilemma that only can be avoided if the mind is divisible. The trouble is that it is neither clear that the mind needs to be divisible in order to avoid this dilemma, nor in fact that there *is* a dilemma here at all.

3

The conceptual argument is not the only argument that have been used to support the partitioning claim. Another is what I shall call the 'explanatory argument'. According to the explanatory argument we need to assume that the mind is divided to *causally explain* the occurrence of weakness of the will. This argument is different from the conceptual argument in that the latter did not assume that the partitioning claim must enter into the causal explanation of weak-willed action; it only held that what makes weakness possible is that the mind is divided into parts. Let me begin by distinguishing two versions of the explanatory argument, depending on the view of mental parts presupposed.

One version argues that the partitioning claim enters into the explanation of weakness of the will *via* the notions of *mental distance* and *cohesion*, in combination with the claim that mental parts have a sustained temporal existence. This version of the explanatory argument is based on the (b)-interpretation of the partitioning claim.[18] The following provides an illustration of how this is meant to work: suppose I accept another drink at the party, despite having judged that it would be better to abstain, and let us assume that one goal I have is to seek pleasure whenever I can. According to the (b)-theorists, this goal might be the centre of a permanent structure of propositional attitudes that constitutes one part of my mind. Another part of my mind might consist of an achievement-oriented goal-structure. Because of mutual coherence, my desire for another drink will naturally gravitate towards the former structure. Once integrated, the internal cohesion between this desire and the other beliefs and desires in the

structure, causes it to become even more deeply entrenched, and also more *distant* from the beliefs and desire of my achievement-oriented goal-structure. But, by becoming more deeply entrenched in this way, it also becomes *more causally efficacious* than it otherwise would have been. Thus, it draws causal strength from the other items in the structure. According to the (b)-theorists, this fact contributes towards explaining how this desire can cause the formation of the incontinent intention. The partitioning claim thereby enters into the explanation of weakness of the will.

Now if holism about the mental is correct, it seems plausible that the mind consists of (more or less) permanent clusters of internally cohesive propositional attitudes, as is suggested by the (b)-theorists. It also seems plausible that any atomic mental item will become drawn into one of these clusters and, as a consequence, may become causally reinforced. The question is whether this feature sheds any light on weakness of the will *specifically*. Clearly, there may be all sorts of general features of the mind which, although they are part of what it is to have a mind, do not necessarily contribute particularly towards the explanation of weakness of the will. The trouble is that if mental parts are permanent structures of the mind in the sense suggested by the (b)-theorists and no mental items are allowed to stay atomic (for holistic reasons), then *all* human behaviour, it seems, whether rational or irrational, must issue from particular mental parts. But then the partitioning claim cannot explain why, *in a case of weakness*, an agent's incontinent desire becomes more causally efficacious than her desire to exercise her rational judgement. What needs to be explained is not the *particular* causal efficacy of her incontinent desire, but its *relative* causal efficacy compared with her desire to exercise her rational judgement. The (b)-theorists' version of the partitioning claim does not seem to contain any resources to explain the latter. It remains unclear, therefore, how it can justify mental partitioning in cases of weakness.

But talk of 'distance' and 'cohesion' may not be the only way of extracting explanatory resources from the partitioning claim. The (c)-theorists proposes an alternative approach. According to their proposal, the partitioning claim enters into the explanation of weakness *via* the notion of mental parts as *rational centres of agency*. This explanation is supposed to work in the following way: the most natural example of a rational centre of agency is a person. The explanatory resources of the view that a person is a rational centre of agency lies in the fact that a person can make things happen *because* of the beliefs and desires she has. By analogy, if we assume that mental parts are rational centres of agency, the explanatory resources of the partitioning claim can in a similar way be located to the fact that *mental parts*, like persons, can make things happen *because* of the beliefs and desires *they* have. Now according to the (c)-theorists, what needs to be explained in weakness is not simply the irrational causation of the incontinent intention or action, but why the agent does not do anything to prevent this causation from occurring. On the assumption that irrational parts are rational centres of agency, there is a simple explanation for this: the irrational part has *a reason* for preventing the cautionary belief from intervening in the rational part,

namely the desire to make the agent perform the incontinent action. To make the rational part form the incontinent intention, it therefore prevents the cautionary belief from having its normal rational effects in the rational part.[19]

It is fair to say, I think, that few have accepted the idea that mental parts are rational centres of agency. Against this idea, it has been objected that it leads to an intolerable regress of mental parts, i.e. a replication of rational parts within irrational parts (Gardner) or (equally bad) a replication of irrational parts within rational parts (Johnston); in neither case does it solve the problem of accounting for weakness but only reproduces the difficulties at a deeper level.[20] I shall not discuss any of these familiar objections here, which anyway seem pretty plausible to me. Instead let me mention another objection which is that the (c)-interpretation is simply very unclear. For what exactly is the relation between the incontinent desire and the irrational part supposed to be on this view? The (c)-theorists claim that the irrational part has a reason for preventing the cautionary belief from intervening in the rational part, namely the desire to make the agent perform the incontinent action. But if this is the irrational part's *reason* and this reason explains why it prevents the cautionary belief from intervening, how can the irrational part be individuated by what appears to be *an effect* of this very act, namely the failure of the cautionary belief to intervene? Surely, the desire to make the agent perform the incontinent action cannot be the irrational part's reason *before* the irrational part has come into existence?

Now both the (b)-theorists and the (c)-theorists reject the (a)-interpretation of the partitioning claim on the grounds that no explanatory resources can be extracted from it. The trouble with this interpretation, they argue, is that it claims that the production of an irrational effect by a mental cause is what defines the boundary of a mental part, at the same time as it holds that weakness of the will *is* the production of an irrational effect by a mental cause. The trouble is that it then follows from *the definition* of weakness that the mind is separated into mental parts in cases of weakness. Introducing semi-independent mental parts cannot be explanatory in that case, only redescriptive. In order for it to be explanatory, the parts must be able to play an active role in the causation of the agent's irrationality.[21]

Although this criticism is correct to point out that no explanatory resources can be extracted from the partitioning claim on the (a)-interpretation it is, I think, somewhat misguided. The point of the (a)-theorists, as I understand it, is that the incontinent agent is controlled by only a part of her mind, which is supposed to be a claim about the sort of *fragmented agency* involved in cases of weakness. For (a)-theorists, the explanatory question is how *this* fact can be explained, not how this fact *can explain* the agent's irrationality. Naturally, the former explanation must coincide with the explanation of the agent's irrationality for the simple reason that the agent's irrationality *is* the fact that she is controlled by only a part of her mind. It will not, however, *include* the partitioning claim.

This being said, it is easy to see how the (a)-interpretation could be expanded in a way that would accommodate the (b)-version of the explanatory argument. For even if one accepts that the individuation of mental parts *in cases of weakness*

necessarily involves a breakdown of reasons-relations, one is not committed to the view that there are *no* mental parts in the mind *independently* of weakness. For example, it is perfectly consistent with the (a)-interpretation to hold that *some* mental parts are permanent features of the mind and should be individuated by mental distance and cohesion. Furthermore, such mental parts could be imagined to enter into the explanation of the breakdown of reasons-relations in ways similar to how such mental parts are supposed to enter into the explanation of weakness according to the (b)-theorists. Thus, they would explain how the incontinent desire evolves into becoming a semi-independent mental part, and thereby also indirectly explain the agent's irrationality.

However, as I have argued, the (b)-theorists' version of the explanatory argument fails to show that the notions of mental distance and cohesion do any specific explanatory work in cases of weakness. So where does this leave the (a)-theorists? Their trouble is that once we give up the conceptual argument, and there in addition are no explanatory advantages of introducing mental parts on the horizon, it is unclear what would motivate talk about separate mental parts at all. What is the difference between saying, on the one hand, that the incontinent agent is controlled by only a part of her mind rather than her whole mind, and saying, on the other, that she is controlled by her strongest motivation rather than by her better-judgement? One is free, of course, to call the various beliefs and desires which make up the agent's strongest motivation, *a part of her mind*. However, talk about 'parts' does not add anything to talk about the agent's motivation as sometimes coming apart from her evaluation. In fact, talk about 'parts' only tends to lead to metaphysical confusion and nothing seems to be lost if we drop it altogether.

Conclusion

In summary, there appear to be neither conceptual nor explanatory reasons for accepting the partitioning claim. So what can be said about the explanation of weakness of the will if this claim is abandoned? In my view, very little in terms of a general account. What we want to explain is the incontinent agent's failure to integrate her evaluative judgement into her motivational structure; specifically we want to explain the relative causal efficacy of her incontinent desire compared with her desire to exercise her rational judgement. Given the psychological complexity of the phenomenon, the possible variety of individual cases and so on, it would be a mistake to rule out the possibility that different explanatory mechanisms may play a role, either alone, or in combination. Let me end this paper by briefly mentioning some familiar mechanisms which together, I believe, explain most individual cases of weakness of the will (although I do not rule out that other mechanisms may play a role as well). None of these mechanisms depend on any claims about mental division.

(i) One familiar mechanism is the agent's perception of *salience*.[22] The incontinent agent is typically *struck* by one particular feature of one of her

alternatives, with the result that the strength of her desire for that alternative increases. This may happen without affecting her most rational judgement that this alternative is inferior in value to her other alternatives. The salient feature may dominate her visual field or perhaps just trigger some strong appetite in her. For example, she may judge that she should abstain from another drink, yet her desire for a glass of wine might draw her attention to a certain attractive feature of wine, such as its calming effect, which increases the strength of her desire for another drink, in turn causing her to have a glass of wine against her own rational judgement that she should not. Many cases of weakness of will can undoubtedly be explained on the basis of such mechanisms.[23]

(ii) Another mechanism that may explain some cases of weakness is our so-called *bias towards the near*.[24] Sometimes we tend to prefer a smaller but nearer reward to a larger but more distant. One way of putting this is to say that we have a discount rate in respect to time, and that we discount the nearer future at a greater rate than the further future. Consider the following example. Suppose that at t_1 I judge that it is better, at t_2, to go early to bed and feel rested at work tomorrow than to stay up late and feel tired at work tomorrow. Suppose nothing changes in my situation between t_1 and t_2 other than my perception of time. Still, just before t_2, I change my judgement and *now* judge that it is better to stay up late than to go early to bed. This is an example of an irrational reversal of preference. A similar mechanism may be at work in some cases of weakness of the will; the weak agent judges it better to do A than to do B, but due to the proximity of the rewards of doing B, her desire to do B is stronger than her desire to do A, so she does B. One reason for some caution at this point may be that it is not entirely clear whether agents who suffer preference reversals of this kind can be said to act *against* their own judgement that it would be better to do something else, or whether they simply change their mind. If the latter is the case, their preference reversals may be irrational, but need not be weak-willed. Another reason for caution may be that not all cases of weakness exhibit the structure necessary for time discounting, i.e. in some cases of weakness proximity of rewards does not seem to play any role.[25] However, I do not see any a priori reason to rule out that *some* cases of weakness conform to this structure, and where, in addition, the agents *do* act against their own judgement that it would be better to do something else. In such cases, time discounting mechanisms may explain the occurrence of weakness.

(iii) A third mechanism that may explain yet some other cases of weakness may be cognitive dissonance, which has been extensively studied in psychology.[26] Suppose we all have a general desire to reduce mental conflicts in order to eliminate psychic tension or discomfort, either by adjusting our behaviour to our belief-system or by adjusting our belief-system to our behaviour. The long-time smoker who is being offered a cigarette but tries to quit smoking, may experience an intensly unpleasent feeling associated with mental conflict. Even if she judges it would be better to abstain all things considered, she may accept the cigarette just to reduce the unpleasent feeling of mental tension, telling herself (perhaps) that one single cigarette cannot make any difference and that this will be the last,

thereby attempting to adjust her belief-system to her behaviour. She may do this, at the same time as she realizes the hollowness of her own justification.

This list of explanatory mechanisms may be far from complete, but do cover, I think, the most typical cases of weakness. Sometimes they will be sufficient alone, other times they will combine, i.e. time-discounting is probably often connected with the perception of salience. What they have in common is that they explain the relative causal efficacy of the agent's incontinent desire relative to her judgement of what it would be better to do, without any appeal to a claim about mental partitioning.[27]

Edmund Henden
Department of Philosophy
University of Oslo
P.O. Box 1020
Blindern, 0315 Oslo
Norway
edmund.henden@filosofi.uio.no

NOTES

[1] Davidson 1982.
[2] Davidson 1982: 304.
[3] Gardner 1993.
[4] Gardner 1993: 61.
[5] Gardner 1993: 62.
[6] Pears 1984.
[7] This concept of the 'cautionary belief' is borrowed from Pears. See Pears 1984: 69.
[8] Davidson 1982: 297-304.
[9] By 'full-blown intentional action' I mean free, deliberate, purposive actions.
[10] Although some seem to believe (wrongly, I think) that it is this general version of the claim Davidson has in mind. For an example, see Mele 1987: 78.
[11] Davidson 1982: 305.
[12] Davidson 1982: 299.
[13] McDowell 1996: xv.
[14] Davidson describes the dilemma in the following passage: '[…] we face the following dilemma: if we think of the cause in a neutral mode, disregarding its mental status as a belief or other attitude – if we think of it merely as a force that works on the mind without being identified as part of it – then we fail to explain, or even describe irrationality. Blind forces are in the category of the non-rational, not the irrational. So, we introduce a mental description of the cause, which thus makes it a candidate for being a reason. But we still remain outside the only clear pattern of explanation that applies to the mental, for that pattern demands that the cause be more than a candidate for being a reason; it must *be* a reason, which in the present case it cannot be. For an explanation of a mental effect we need a mental cause that is also a reason for this effect, but, if we have it, the effect cannot be a case of irrationality. Or so it seems' (Davidson 1982: 299).
[15] Davidson 1982: 300.

[16] Of course, *the fact* that I have this desire could be your reason for entering my garden if you wanted to satisfy my desire.

[17] Perhaps Davidson's response to this objection would be to reject the distinction between normative and motivating reasons on the grounds that it is ruled out by his general interpretationist view of the mental. My reply would be that if this is the case, it should be more of a worry for the interpretationist than for the proponent of a distinction between normative and motivating reasons. The latter distinction appears to be of deep importance for the understanding of many difficult issues in moral psychology. However, I am not convinced that the interpretationist needs to give up the distinction between normative and motivating reasons. For an attempt to use this distinction within a broadly Davidsonian approach to weakness of will, see my 'Intentions, All-out Evaluations and Weakness of the Will' (forthcoming in *Erkenntnis*).

[18] For an example, see Gardner, 1993: 63. My illustration below is based on one of Gardner's own examples.

[19] This is how I interpret Pears' view in *Motivated Irrationality*. See especially p. 87.

[20] For Gardner's criticism of the (c)-interpretation, see Gardner 1994: 73-74. For Johnston's criticism of this interpretation, see Johnston 1995.

[21] See Pears 1984: 84 and Gardner 1994: 61.

[22] The perception of salience has been mentioned by many authors in connection with weakness of the will. See for example Rorty 1980: 193-212, Mele 1987: 84-93, Pears, 1984: 174-179.

[23] One question that may arise is whether something like the (b)-theorist's conception may not after all return on the explanatory scene at this point. For will not the *onset* or *activation* of one particular salient feature rather than another, *itself* demand an explanation, which can only be given in terms of the agent's more permanent goal-structures? It may well be true that there are cases in which one of the agent's permanent goal-structures do contribute towards explaining the coming into play of a certain salient feature. However, I do not see any reason why this should be true in *all* or even *most* cases of weakness of the will. Even a person with a strong achievement-oriented goal-structure and a very weak pleasure-oriented goal-structure may have her attention drawn to a certain pleasurable feature of one of her alternatives, having her desire for that alternative reinforced as a result. In general, the explanation of what features of alternatives strikes us as being salient seems to have more to do with our common human nature than what rational goal-structures happen to be part of our individual psychological makeups.

[24] Examples of authors who mention this mechanism in connection with weakness of the will are Elster 1999, Gjelsvik 2000.

[25] Mele discusses such cases. See Mele 1987: 86.

[26] The classic work on cognitive dissonance is Festinger 1957.

[27] I would like to thank David Charles and Bill Child for helpful comments on earlier drafts of this paper.

REFERENCES

Davidson, D. (1982), 'Paradoxes of Irrationality', *Philosophical Essays on Freud*, eds. R. Wollheim and J. Hopkins, Cambridge: Cambridge University Press, pp. 289–305.

Elster, J. (1999), 'Davidson on Weakness of the Will and Self-Deception', *The Philosophy of Donald Davidson*, ed. Lewis Edwin Hahn, LaSalle, Illinois: Open Court, pp. 425–440.

Festinger, L. (1957), *The Theory of Cognitive Dissonance*, Stanford, CA: Stanford University Press.

Gardner, S. (1993), *Irrationality and the Philosophy of Psychoanalysis*, Cambridge: Cambridge University Press.

Gjelsvik, O. (2000), 'The Epistemology of Decision-Making Naturalised', *Knowledge, Language and Logic: Questions for Quine*, eds. Alex Orenstein and Petr Kotatko, Dordrecht: Kluwer Academic Publishers.

Henden, E. (2004), 'Intentions, All-Out Evaluations and Weakness of the Will', forthcoming in *Erkenntnis*.

Johnston, M. (1995), 'Self-Deception and the Nature of Mind', *Philosophy of Psychology, Debates on Psychological Explanation*, ed. Cynthia Macdonald, Oxford: Blackwell, pp. 63–91.

McDowell, J. (1996), *Mind and World*, Cambridge, Mass: Harvard University Press.

Mele, A. (1987), *Irrationality, An Essay on Akrasia, Self-Deception, and Self-Control*, Oxford: Oxford University Press.

Pears, D. (1984), *Motivated Irrationality*, Oxford: Oxford University Press.

Rorty, A. (1980), 'Akrasia and Conflict', *Inquiry*, 22, pp. 193–212.