Philosophy and Phenomenological Research
Vol. LXVII, No. 1, July 2003

Self-Deception, Interpretation and Consciousness

PAUL NOORDHOF
University of Nottingham

I argue that the extant theories of self-deception face a counterexample which shows the essential role of instability in the face of attentive consciousness in characterising self-deception. I argue further that this poses a challenge to the interpretist approach to the mental. I consider two revisions of the interpretist approach which might be thought to deal with this challenge and outline why they are unsuccessful. The discussion reveals a more general difficulty for Interpretism. Principles of reasoning—in particular, the requirement of total evidence—are given a weight in attentive consciousness which does not correspond to our reflective judgement of their weight. Successful interpretation does not involve ascribing beliefs and desires by reference to what a subject ought to believe and desire, contrary to what Interpretists suggest.

Have you ever believed something you feel to be quite impossible to give up while consciously believing that the evidence makes it more likely that what you believe is false rather than true? I shall argue that such cases are possible. They throw into doubt standard analyses of self-deception. They are not cases of self-deception and yet, given some background assumptions and a bit of development, they would be classified as self-deception by the extant analyses in the literature.

They also throw down a challenge to Interpretist approaches to the mental. The Interpretist holds that we may learn all there is to know about the nature of beliefs, desires and other propositional attitudes by considering the role they play in interpreting agents' behaviour (Dennett (1981a), p. 15; Davidson (1983), p. 315; Child (1994), pp. 1, 24, 47-55). Beliefs and desires are the characteristic components of an interpretive scheme thrown over an agent's activities governed by the norms of rationality and the good by which an agent, largely, has to abide if propositional attitudes can be ascribed to him or her. If an agent is to be interpretable at all, he or she should normally be rational and a lover of the good (Davidson (1970), p. 222). Predictions about how an agent will behave are derived from the attributions of belief and desire by considering what it would be rational to do in the light of those beliefs and desires. Hence, what it is to be a true believer and desirer is to be 'a sys-

tem whose behavior is reliably and voluminously predictable' by these means (Dennett (1981a), p. 15).

Although belief in the face of conscious evidence to the contrary fails to count as self-deception, such cases involve irrationality if something like the following is a plausible principle of belief formation.

> Give credence to the hypothesis most highly supported by all available relevant evidence (Davidson (1985), p. 140).[1]

Following Donald Davidson, I will call this the *requirement of total evidence*. Davidson takes this principle of belief formation to be partly *constitutive* of rationality and hence the ascription of beliefs. If this is right, then for agents to be interpretable, they must be taken to 'embrace' the requirement of total evidence (Davidson (1985), pp. 141-142, 147). I will suppose that, at the minimum, in the kind of case I am about to outline, the agents subscribe to the principle and so are irrational to depart from it. The cases I describe show the important role that an appeal to conscious attention plays in characterising self-deception. I shall argue that Interpretists cannot capture the nature of this appeal.

My discussion proceeds as follows. First I give a description of an instance of the kind of case I have in mind and compare it to others. Then I explain why such cases pose a challenge for standard analyses of self-deception. Finally, I discuss Interpretism and isolate the challenge that self-deception presents it.

1. The Faithful Lover

Probably most of us have experienced a situation in which there are clear signs that a relationship has broken down and yet one of the parties insists that the other still loves him or her and that the relationship is not really over. Sometimes the person who clings on to the relationship—call him 'Fido'—seems to be in the following state of mind.

> [A] *Fido is attentively conscious of the fact that the evidence shows that it is more likely that she does not love him any more than that she does and, hence, believes that the evidence shows that it is more likely that she does not love him any more than that she does.*

If a subject is attentively conscious of something then he or she is focusing on it and not distracted by other things. If a subject has an understanding of what he or she is attending to, then the subject will be aware of its most significant manifest features. Fido is aware of the evidence that she does not love

---

[1] Davidson credits Carnap and Hempel for the name and the formulation of the principle (see Carnap (1950), pp. 211-213; Hempel (1965), pp. 397-403).

him any more. She does not return Fido's calls. She tells Fido that she is going out with somebody else and doesn't love him any more. Their mutual friends, somewhat unkindly, tell Fido she disparages him behind his back. When she does meet him, she seems to treat him with indifference.

Fido recognises that the evidence strongly favours the claim that she does not love him any more. He does not have an idiosyncratic notion of evidence. He has shown in the past that he holds that evidence of this strength would justify a belief that she does not love him any more. He encouraged others to draw the same conclusion in similar circumstances that occurred to them. Moreover, he acknowledges that his present circumstances are relevantly similar to those in which he counselled others to form a different belief. He seems sincere in his acknowledgement. Nevertheless, as he puts it, he sees now how they could not stop believing that their lover still loved them if they felt the way he feels now.

Fido's awareness of the evidence and grasp of its significance makes the ascription of the belief, that the evidence shows that it is more likely that she does not love him any more than that she does, plausible. His awareness of the evidence and recognition of what it indicates makes it legitimate to suppose that this belief is conscious.

[B] *Fido is attentively conscious of his belief that she still loves him.*

Fido says with the utmost earnestness that, in spite of all, she still loves him. This, by itself, is not enough to make the attribution of the belief credible, although it may make one pause in attributing the opposite belief. However, there are other features of Fido's state that add to the impression that he believes that she still loves him. First, it is clear that Fido does not feel the distress that we would expect him to feel if he really did believe that she does not love him. Of course, there might be other explanations for the absence of distress. But if we find that later on Fido is distressed and he says that he now realises that she doesn't love him any more, it would be very plausible to attribute to Fido the prior belief that she still loved him. We would say that, although he appreciated the evidence that she didn't love him, it hadn't really sunk in to form the belief.

Second, Fido claims that it is impossible for him to believe that she does not love him any more and we begin to see how the belief is supported by some very strong attitudes. They seem to originate in the early days of his relationship with her. According to Fido, the relationship began intensely. They felt very close. During the course of this they had discussed how misunderstandings might arise as a result of which people drift apart. They had promised each other not to forget the original feeling of closeness they had and to work to overcome any misunderstandings which might occur. She had

confessed that she could be cold and distant at times and forget what had gone before. 'Don't give up on me,' she had said. Fido has resolved not to.

One way to seek to characterise Fido's recollection of the early days of the relationship is that he has evidence that she still loves him. I am prepared to concede this up to a point. I think that these recollections may count as evidence that it is *possible* that she still loves him. Nevertheless, this does not imply that it is incorrect to ascribe to Fido the belief that the evidence *in toto* shows it is more likely that she does not love him than that she does. Fido knows it is implausible that the recollections do outweigh the evidence that she does not love him. He is well aware that relationships are littered with the kinds of exchanges he mentions without it being true that one of the parties still loves the other while giving every impression to the contrary.

Instead, Fido's recollection of these exchanges has given rise, in Fido, to strong feelings of faithfulness and loyalty to the relationship. This is revealed by the kind of recollections that play a central role in Fido's mental life. If Fido focused on the evidence that she could be cold and distant and yet, in fact, turned out to care, then perhaps these recollections would act mainly as reassurance and reduce anxiety about the counter-evidential belief. This would depend upon him viewing as, at least, questionable the claim that the evidence in toto shows that it is more likely that she does not love him than that she does. Instead, suppose that the focus is on the emotional temperature of their exchange when she emphasised how close she felt to him, about their not betraying each other but persevering with the relationship, and all the other things that might veer into sentimentality. Then it seems much more plausible that the recollections support the belief that she still loves him by working Fido up into a state of loyalty and faithfulness. In such a state, it seems to him psychologically impossible to doubt her continued love for him however difficult she is currently being. When he thinks about how the evidence favours the belief that she does not love him, it just strikes him that it would be wrong for him to cease to believe in her love. He would be giving up on something important and valuable. The belief that she still loves him sustains him in his attempt to resuscitate the relationship. He recognises that the requirement of total evidence recommends that he abandons his belief that she loves him. However, this strikes him as an inappropriately cold and dispassionate way of looking at things. In effect, the requirement no longer overrides all other considerations in his reflection.

It is tempting to suggest that if Fido really believes that she still loves him he must be taking the evidence quotationally. He recognises that other people would call the evidence 'evidence which shows that it is more likely that she does not love him than that she does' but it does not seem that way to him now. This mirrors what some people are inclined to say if it is suggested that I may believe that I ought to do A and yet be motivated to do B

instead. However, the same response would seem to be available in the case of beliefs about evidence as is available in the case of moral beliefs. In the moral case, we can allow that although there is an internal connection between our moral beliefs and our desires, it is not invariable. It just holds if we are practically rational. Thus

> If S believes that it is right for him or her to do A in circumstances C and S is practically rational, then S is motivated to do A in C (Korsgaard (1986), pp. 8-9; Smith (1994), pp. 61-63).

We don't have to write off all failures of motivation as implying that we only believed that A was what other people called 'right'. It may be right by our own standards but we are practically irrational. Similarly, we can allow that although there is an internal connection *between* the belief that the evidence shows that it is more likely that p is true than that not-p is *and* believing p, the connection is not invariable. We may record the connection as follows.

> If S believes that the evidence shows that it is more likely that p is true than that not p is and S is (theoretically) rational, then S will believe that p.

I do not claim that Fido is theoretically rational. His failure to form beliefs in line with the requirement of total evidence shows otherwise.

Given that the analogy between theoretical reason and practical reason holds, there is room for cases of the kind I have suggested that Fido exemplifies. Moreover, some of Fido's behaviour reinforces the need for allowing that such cases are theoretically possible. He does not look upon the evidence that it is more likely that she does not love him than that she does with equanimity. The evidence does not just match other peoples standards but accords with his own. For that reason, he rehearses 'cover' stories which explain how it could still be true that she loves him in spite of the evidence. For instance, he argues 'She is frightened of becoming too close. She is shy. Her jaunty appearance and extrovert personality are a cover for this shyness', 'She is fighting against her love for me so that she can retain her independence' and so on.

Fido's rehearsal of a cover story is not endorsement of a positive interpretation of the evidence, turning it into evidence that she loves him after all (cf. Szabados (1973), p. 205; Mele (1997), p. 94). Fido recognises that the most plausible interpretation of the evidence is that she does not love him. However, he is convinced that she does love him and is reaching for any story which reduces his unease at believing something in the face of the evidence. He exploits the fact that empirical evidence never entails the truth of a belief

about the world around us and entertains stories which play upon his emotions further to bolster his belief that she still loves him.

The cover stories therefore have two roles. The focus of his recollections make the favoured belief very attractive to him due to the loyalty and faithfulness they fan. One role of the cover stories is to reduce the anxiety resulting from the conflict between the favoured belief and the evidence that it is more likely that she does not love him than that she does. It is not that Fido comes to the view that the evidence is misleading. Nor does this follow from the fact that he believes that she still loves him. By hypothesis, he is not theoretically rational. The conflict upon which the irrationality is based is genuine. So Fido both has the belief and remains of the view that the evidence in toto favours the opposite conclusion. Rather the cover stories emphasise the *possibility* that the evidence might be misleading and it is this that allows for the attractiveness of the belief to win out in the way I identified earlier.

The second role of the cover stories stems from the focus of the recollections. Talk of shyness, being frightened of becoming too close, and dwelling on features of her personality that made her attractive to him, as well as (perhaps) frustrating, all have an emotional dimension. The style of the talk is important. He is not referring to the results of statistical studies on the dynamics of a relationship and such like. He is telling himself the stories in language that engages the emotions. The cover stories further quicken the emotions that support and bolster his belief that she still loves him.

The role of the cover stories in Fido's case is also brought out by comparing it with standard cases of partisanship. In partisanship, a subject's commitment to the truth of a certain proposition can be the basis for identifying evidence in favour of it (Morton (1988), pp. 176-178). Examples of partisanship include research scientists (or philosophers) convinced by a certain theory and adherents to a particular political viewpoint. The case of Fido seems importantly different. The entertaining of various cover stories are not the basis of attempts to find evidence for the truth of the belief that she still loves him. Fido does not treat them with that degree of intellectual seriousness. Instead, the stories reinforce the attitudes that support the belief that she still loves him. He goes from one story to another depending on which strikes an emotional chord.

Some might take the case of Fido to demonstrate that it can never be theoretically irrational to believe that p when, given the rest of one's beliefs, there is some probability that p is true (however low). If that were right, the requirement of total evidence would be incorrect. However, this manoeuvre appears misguided. One result would be that few, if any, beliefs about the external world would count as irrational, even the extreme beliefs of a paranoid schizophrenic.

The case of Fido contrasts with cases of trust. Suppose you are standing in the middle of a circle of people who are all part of a drama course and (as instructed) you let yourself fall, to be caught by them. You trust them to catch you because that is part of the exercise (Holton (1994), p. 63). Do you believe that they will? Perhaps not. There is a hint that they might make a mistake or play a trick on you. So trust seems distinct from belief. One manifestation of this is that one can decide to trust something but not believe it (Holton (1994), pp. 63-64, 69).

Cases of trust share some features with that of Fido. We trust people rather than things and its disappointment is linked to reactive attitudes like a sense of betrayal (Holton (1994), pp. 65-66). Fido believes that *she* still loves him and when, finally, he believes that she does not love him any more, he will feel a sense of betrayal. Nevertheless, there are important differences. First, Fido's belief that she still loves him is not a matter of decision in the way that trust can be. Even if one is prepared to allow that some beliefs can be formed at will, Fido's belief is not one of them. It is psychologically impossible for Fido to abandon this belief and he did not choose to have it to begin with (see Noordhof (2001) for further discussion of belief and the will). Second, Fido would feel more upset if he were just trusting that she loved him rather than believing that she did. Acting as if he believes that she still loves him without believing it would not remove the psychological anguish which would otherwise arise from believing that the evidence shows that it is more likely that she does not love him than that she does. Third, one usually trusts someone to do or fail to do something. Yet, plausibly, love or failing to love someone is not something that we do. It is possible that Fido is sufficiently misguided as to suppose that love or failing to love is something that a person does for which they can be held responsible. Nevertheless, this feature of trust at least throws open to question its application to the present case.

The first two reasons against taking the case of Fido as a case of trust also seem to hold against the idea that he doesn't believe that she still loves him but merely thinks that she does. Thinking in the sense of just entertaining the proposition that she still loves him does not bring with it the kind of (re)assurance that belief would and which is characteristic of Fido's state. Equally, thinking that she loves him can be a matter of decision. At first glance, things might seem better if we took Fido to be merely sincerely avowing that she still loves him. A sincere avowal that she still loves him is likely to be accompanied by a measure of reassurance. Similarly, one can't just decide to avow *sincerely*. However, if sincere avowal is to be distinguished from belief, then that is because one can sincerely avow something on the basis of what one supposes to be the case and yet it will not quite sink in or take hold so as to be an expression of belief (see Audi (1989), pp. 212,

214). A sincere avowal that p would normally be based upon one's appreciation that the evidence shows that p is more likely to be true than not p. The gap between sincere avowal and belief would rest on the failure of the evidence for one reason or another to yield the appropriate belief. But Fido recognises that the evidence points in the *opposite* direction. Therefore, an appreciation of evidence for the proposition that she loves him can't be what makes his avowal sincere. The only other plausible basis of Fido's sincerity is that he actually does believe that she loves him. So we cannot characterise Fido as merely sincerely avowing rather than believing that she still loves him.

In fact, Fido shares the features of other, perhaps more standard, cases of faith. Robert Merrihew Adams has characterised faith as involving 'believing something a rational person might seriously be tempted to doubt' which typically 'includes doubt, and a certain sensitivity to opposing reasons, as well as a certain resistance to them' (Adams (1995), pp. 75, 85). Although faith is resistant to evidence, it still involves the idea of trying to reflect something which is part of reality. The resistance to evidence does not stem from an indifference as to the way reality is. It is just that the loss of the particular belief constitutive of faith is far more significant to an agent than the danger of believing on insufficient evidence. Something of value would be gone from the agent's life. What matters is that the agent believes that p —so long as p is true—and not that the agent reduces to reasonable proportions his or her chance of being wrong about p (Adams (1995), pp. 83-88). As an example, Adams cites having faith that a person's life is worth living even in the face of strong evidence to the contrary such as physical suffering due to a terminal illness (Adams (1995), p. 79). There are other plausible cases. A loving father believes that his son is innocent of a terrible crime out of the type of loyalty that is typical of some parental love. However, he believes that the evidence shows that it is more likely that his son is guilty than that he is not. Of course, he thinks his prior acquaintance with his son provides counterevidence but he recognises that most parents think their child incapable of a serious crime. He does not believe that there is strong enough counterevidence since the evidence against his son is pretty damning. According to some thinkers, religious faith also displays the features of the case of Fido. Indeed, they say it must display them. Thus Sören Kierkegaard writes

> For whose sake is it that the proof is sought? Faith does not need it; aye, it must even regard the proof as its enemy … when faith … begins to lose its passion, when faith begins to cease to be faith, then a proof becomes necessary so as to command respect from the side of unbelief (Kierkegaard (1846), p. 31, see also, pp. 30, 218-221, 333-334, 384, 540).

and Miguel de Unamuno

> And not only do we not believe with reason, nor yet above and below reason, but we believe against reason. Religious faith, it must be repeated yet again, is not only irrational, it is contra-rational (Unamuno (1921), p. 198).

According to both, religious faith in God must involve a belief in the face of a clear perception that the evidence shows that it is more likely that there is no God than that there is. Even if both Kierkegaard and Unamuno are wrong about how religious faith must be characterised, they do demonstrate by their writing the psychological reality of the type of case I have described. The fact that the impression of psychological reality persists over a range of different cases, and related cases such as that of partisanship, suggests that it is legitimate to ascribe the combination of beliefs which I have taken to distinguish the case of the faithful lover.

2. What do cases of this type show about the character of self-deception?

I think that it is clear that Fido and related cases are not self-deceived. They can be attentively conscious of having both the belief that p and the belief that the evidence shows that not-p is more likely to be true than p. Moreover, when they are not, this is not because they are trying to avoid one of the beliefs or hide something from themselves. Nor is it an essential feature of the cases that the beliefs in question are false however perverse they seem. She *may* still love him. So this connotation of deception is not present and cannot serve to motivate a claim that they involve self-deception (see Mele (1987b), p. 135).

If my assessment of this type of case is right, then certain putative analyses of self-deception are inadequate. Some have tried to characterise self-deception simply in terms of a belief that not-p in the face of strong evidence to the contrary (Canfield and Gustavson (1962), pp. 34-36). Some have added to this analysis the requirement that the subject believes that there is strong evidence against the belief that not-p (Penelhum (1964), p. 88). Others still have added that there should be a desire-initiated 'inappropriate' treatment of the evidence in order to generate or support the belief that not-p (Mele (1987a), p. 10; Mele (1987b), p. 127). Fido appears to fit all of these analyses and yet is not a case of self-deception.

The only doubt that may arise concerns the third. It might be wondered whether Fido has the desire that she still loves him. The first thing to point out is that in this context 'desire' is meant to be understood broadly as a pro-attitude of some type or other. So we should not worry about whether desire is the right way to characterise part of what is involved in loyalty and faithfulness. The issue is whether Fido can be ascribed a pro-attitude to the putative fact that she still loves him. It seems that the answer is 'yes'. Retaining faith in the relationship involves there being certain things that one is in

favour of believing and other things that one is not in favour of believing. In the case of Fido, one of the things that his faith makes him in favour of believing is that she still loves him. If somebody has faith, the state of faith itself is represented as attractive. Central to their faith is a belief. In the case of Fido, the belief concerns the continued existence of the relationship and, in particular, the fact that she still loves him. The relationship would not exist if she did not. So Fido desires to have this belief.

It would be a mistake to limit the desire to this, though. Having faith is not being in favour of believing something false. It is being in favour of believing something which, in fact, is true. The value of the belief is diminished if it turns out to be false. The element of faith is that the belief is held in the face of reason not truth. So the faithful person is also in favour of the world being a certain way. Indeed, the pro-attitude towards the world being a certain way is primary. It is the attractiveness of the world being that way that makes belief in it attractive. It is an expression of the commitment of the faithful that they are both in favour of the world being a certain way and place the greatest importance on believing that it is. In the case of Fido, the importance is reflected in the fact that he found it psychologically impossible to give up the belief that she loved him. He would feel that he had lost something of significant personal value: the fact of her love for him. On the basis of these points, I think it is legitimate to ascribe to Fido pro-attitudes to both the belief that she still loves him and, importantly in the case of my conclusion about Mele's work, to the fact that she still loves him (if it is a fact).

There are two ways in which Fido's desire that she still love him supports his belief. The first is obvious. It seems that if Fido were to lose this desire—if Fido is no longer faithful!—then he would no longer believe that he still loves her. This would suggest that Fido's desire causally sustains the belief. Of course, this would leave it open that Fido's desire did not originally cause the belief. However, it is easy to imagine a slight development of Fido's case which would have this implication. Suppose that, for a moment, at the beginning, Fido believed that she did not love him any more. The shock of Fido's fleeting recognition of how things have changed sharpened his feelings of loyalty and faithfulness to the relationship resulting in the picture I described earlier. If he had not received this initial shock, he would have tired of the whole thing, failed to have the desire that she still loves him and, as a consequence, would not now believe that she did. In such circumstances, it seems right to say that the desire that she still loves him caused the belief.

It might be wondered whether Fido could really be attentively conscious of the connection between his belief that she still loves him and his desire that she still loves him. Surely something must be hidden from him? I don't think this is right. We don't lose a belief just through finding out about

inadequacies in the way the belief was formed. What matters is the justification we can find for it now. By the same token, if Fido can have the conscious belief that she still loves him due to the value that he places upon keeping faith with the relationship, then there is no further danger from his being aware that it was his evaluation of the relationship rather than evidence that was behind the formation of the belief in the first place. He might even acknowledge that he was worked up by the original shock. It is just that he would think of it as a piece of good fortune. The shock enabled him to see the value of something he might otherwise, because of various moral weaknesses on his part, have let slip away. This is the positive way of looking at the commonplace wisdom that the fear of loss keeps them keen.

What is missing in Fido and his kin which is present in self-deception? The options appear limited. One suggestion is that the distinctive activity of self-deception—for instance, avoiding evidence—is done with the intention of producing a belief that p as a result of an unwelcome prior and continuing recognition that not-p is true or that the evidence shows that not-p is more likely to be true than p (Davidson (1985), p. 145, see also Talbot (1995), p. 30). There is a question mark over whether self-deception should really be taken to be an intentional action rather than, for instance, a purposeful but non-intentional tropistic mechanism (Johnston (1988), p. 86). However, I do not have to resolve this issue to arrive at a preliminary assessment of the suggestion. To the extent that it makes sense to claim that self-deceivers intend to produce the deceptive beliefs so does Fido and yet Fido is not self-deceived. Fido has a pro-attitude in favour of the belief that she still loves him. By providing various cover stories and recollecting the early days of the relationship, Fido sustains the belief that she still loves him. His provision of cover stories and dwelling on certain recollections are actions rooted in the unwelcome recognition that the evidence shows that it is more likely that she does not love him than that she does and that, because of the requirement of total evidence, he ought rationally to believe that she does not love him. Fido also seems to recognise, by dwelling on the things he does, that he is in some danger of losing the belief if he did not. But these activities are not deceptive. Instead, they support something which Fido clearly sees to be of value. It is because they enable him to see clearly the value of the relationship, support his evaluation of it, and downplay the tension with the requirement of total evidence, that Fido can be aware of them. 'Sure I sometimes worry that I am wrong but I only have to think back on the times we had to *feel* certain that I am not'. Taken together, these make it just as plausible to ascribe to Fido the intention to produce a belief as it does in cases of self-deception where a belief is produced or sustained by avoiding certain evidence. Nevertheless, since Fido may be attentively conscious of what he is up to, he is not self-deceived.

A second suggestion is that self-deception always involves the purposeful reduction of anxiety whereas the cases I have described do not.[2] The first thing to note is that it is far from clear that all cases of self-deception involve the reduction of anxiety. Some might arise from jealousy or anger in a different way. The emotions support respectively a belief about a partner's likely fidelity or the merits of one's antagonist which is against what one believes the evidence to suggest (see Lazar (1999), pp. 280-284; Noordhof (1999), p. 183). It is possible to claim that, in the case of jealousy, someone believes that their partner is unfaithful to reduce anxiety because they cannot stand the prospect of being wrong. In the case of anger, it might be said that we believe that the object of our anger is lacking in merit in order to reduce anxiety in our feeling of aggression towards them. The question is whether all cases of jealous and anger produced belief must be understood in that way. It seems not. Both jealousy and anger involve a pro-attitude towards a certain fact. This may generate a belief in the face of evidence. The belief works the person up further. If jealousy and anger have the purpose of generating certain responses, then beliefs which heighten responses based upon them seem a natural outcome of these purposes whether or not anxiety is also reduced. If we are going to think of self deception in terms of the purposeful production of a belief, then the limitation to anxiety reduction does not seem particularly well-motivated. It seems to mistake what might be true of many cases of self-deception for what must be true. However, to the extent that we are disposed to look at anxiety reduction as distinctive of self-deception, it is not clear that Fido fails to have this characteristic. After all, his belief that she still loves him is purposely sustained and he would be anxious if he lost it. It is just that this does not capture the full dimensions of the case, in particular, the value that Fido places upon the relationship.

A final suggestion to consider is that a self-deceived agent, S, must have a particular set of psychologically disturbing beliefs about his or her recent cognitive history. According to Scott-Kakures they are: (i) S believes that p; (ii) S believes that, at some prior time t, he or she believed that not-p; (iii) S believes that, at t, he or she had sufficient reason for believing that not-p; (iv) S believes that there has been no chain of reasoning to rationalise the transition—there is a fissure (Scott-Kakures (1996), pp. 50-51). However, once more, a slight development of the case of Fido would make it plausible to ascribe these beliefs to him without thereby attributing self-deception. Reflecting on the events of the last few months, Fido believes that, at a certain moment two months ago (call it 't'), he had sufficient reason for believ-

[2] This suggestion has been put forward by Mark Johnston, and endorsed by Annette Barnes, as a distinguishing feature of self-deception. They are no particular friends of the Interpretist approach (Johnston (1988), p. 86; Barnes (1997), p. 117). My aim is just to assess whether it is something to which the Interpretist might appeal.

ing that she did not love him any more. He believed then, as he does now, that the evidence shows that it is not the case that she loves him is more likely to be true than that she does. On further reflection, Fido might believe that, at t, for a moment, he believed that she did not love him any more. That was his darkest hour. Then, recalling all the conversations in the early days of their love, his emotions were engaged and he became convinced that she still loved him. He believes that the transition is not rational. As far as I can see, Fido does not become self-deceived by being aware of the history of his beliefs, nor, with the qualification below, would he be self-deceived if he had not engaged in the reflection I described and hence was unaware of how things had gone (see also Barnes (1997), p. 145, fn. 14).

I have tried to explain how the case of Fido presents a problem for the main theories of self-deception in the literature. Doubtless there will be further refinements of these theories but it is not at all clear how the attribution of further propositional attitudes or the identification of more complex purposes will change the preliminary conclusion we have reached. The case of Fido and his ilk display a resilience to attempts to obtain the right verdict by further accretions of this kind. This is the basis for my claim that the crucial difference between Fido and those who are self-deceived can only be described in terms of attentive consciousness. A preliminary characterisation is that Fido is not self-deceived because he is attentively conscious of the belief that the evidence shows that it is more likely that she does not love him than that she does and attentively conscious of the history leading up to the formation of the belief that she still loves him. There is nothing that he is hiding from himself or merely taking account of without focusing on it (see Fingarette (1998), pp. 294-295). He ruefully acknowledges these facts but still feels convinced that she loves him. However, this does not quite capture what is distinctive of self-deception. A moment's inattentiveness to one of these things would not make Fido, for that moment, self-deceived. The distinctive feature of the self-deceived is that a failure of attentive consciousness enables them to possess or retain a belief that they would not otherwise have. However, Fido can retain the belief that she still loves him even if he appreciates that the evidence points in the other direction. He endorses the belief because of the central role which it plays in his life.

More precisely, then, the crucial difference between the kind of case I have described and cases of self-deception is that self-deception requires the following two conditions to be met.

(a) The subject, S, fails to attend consciously to either the evidence which rationally clashes with the motivationally favoured proposition which he or she believes or some element of the psychological history characteristic of the self-deception behind the belief in the motivationally favoured proposition.

(b) If the subject were to attend consciously to both the motivationally favoured proposition and either the evidence which rationally clashes with it or the psychological history (whichever applied from clause (a)), the motivationally favoured proposition would no longer be believed.

Clause (b) captures the instability inherent in self-deception. The subject is on the verge of having to abandon the motivationally favoured belief but is saved by crucial lacks of attentive consciousness. The instability does not always have to arise from contradicting beliefs. The kind of relation between the beliefs that gives rise to the instability may vary from circumstance to circumstance. The presence of attentive consciousness in some circumstances will have different consequences from those it has in others (cf. Pears (1984, 1986), pp. 73-76)). If you badly want to win the lottery because you are deeply in debt, you may persist in believing that you have a good chance of winning in spite of your conscious attention to what you believe to be strong evidence that the odds are against you ('You just never can tell'). On the other hand, if you want to win the lottery because that would be fun, life is going pretty well as it is, conscious attention to your belief that there is strong evidence that the odds are against you will make you abandon your belief that you have a good chance of winning.

Clause (a) is needed to distinguish between self-deception and a related lack of integration. There are plausible cases in which the subject fails to consciously attend to the motivationally favoured proposition, consciously attending rather to the propositions which rationally clash with it. Yet the subject persists in believing the motivationally favoured proposition. This may be what happens in some cases of shock. A person has just heard, and so is all too attentively conscious of the fact, that he or she has been sacked from a job. Yet he or she does not feel upset. In such circumstances, we talk in terms of it not really sinking in. We might capture this by saying that the person still believes that he or she has the job. He or she has not abandoned this belief because he or she is transfixed by the belief that he or she has just been sacked and has not put the two things together. It may take a little while for the agent to attend sufficiently to the clash between the fact that he or she has been sacked and the fact that he or she has got a job to finally lose the belief that he or she has got a job with the loss of self-esteem and purpose that may provide. This is not self-deception.

The conditions cover cases in which a subject, who believes that p, systematically avoids situations in which he or she might obtain evidence against p, because of a desire to believe that p. I think, intuitively, we would count these as cases of self-deception. The conditions don't just advert to what would happen if a subject were attentively conscious of counterevidence. They also advert to what would happen if a subject were attentively conscious

of an element of the psychological history behind the belief in the motivationally favoured proposition. Part of this psychological history would be the agent's systematic avoidance of the evidence against p. If attentive consciousness to this element of the psychological history undermined the belief that p, then I think we have self-deception. If, on the other hand, attentive consciousness did not undermine the belief, then we have deliberate avoidance of, and indifference to, the evidence but not self-deception. The latter type of case is apt to be less common. It would be peculiar systematically to avoid attending to evidence against the belief that p if the evidence had no tendency to undermine the belief if it were attended to. Nevertheless, we can imagine people doing it, those with an irascible temperament and a closed mind.

I have not provided a full analysis of self-deception. My point is rather that, whatever your view about the beliefs, desires and intentions distinctive of self-deception, in addition we will have to add the clause about attentive consciousness. I have couched the clause vaguely in terms of the psychological history characteristic of self-deception so that it may be plugged into any account of the characteristic states of self-deception. For the purpose of subsequent discussion, what is important is that the need for a clause of this type causes problems for the Interpretist's position. To see this, we must begin by getting a clearer idea of the nature of Interpretism and the explanatory resources to which it appeals.

3. Consequences for the interpretist approach to the mental

As I noted at the outset, the interpretist approach to the mental holds that the nature of beliefs, desires and other propositional attitudes is given by their role in a normative scheme of interpretation of agents' behaviour. As a result, interpretist approaches have to provide an explanation of why attributions of irrational action, irrational belief formation, and inappropriate desire are not prima facie evidence that either we've made the wrong attribution of beliefs and desires or we shouldn't have applied the interpretive scheme at all.

Interpretists have rightly taken self-deception, in particular, as a problem for their approach. One reason for this is that it seems that we can make sense of what is going on in self-deception. Interpretists link making sense of an agent with seeing them as rational and lovers of the good. Yet, self-deception is a form of irrationality. How can we, on the one hand, make sense of the agent and on the other view them as irrational? A second, and related, reason why self-deception presents a problem is that we are fairly settled in our attribution of the characteristic patterns of beliefs and desires constituting self-deception yet it seems that we shouldn't be. Given that the attribution of a self-deceptive pattern involves a departure from rationality it seems that our confidence should be less. After all, there might be a better interpretation which makes the agent more rational. It may be that the constraints which

govern interpretation indicate that the attribution of one of the characteristic self-deceptive combinations of beliefs and desires is the best that we can do but it is not obvious that this is so. We might minimise irrationality by making agents more devious in their interactions with us, yet we don't.

The line adopted by Interpretists (or on their behalf) to deal with self-deception and related kinds of irrationalities is that the agent has divided into two possibly overlapping subsystems in which some of the elements of one are separated from the other by a mental partition. This partition is understood partly in terms of the *absence* of a certain kind of rational connection between the relevant beliefs, in this case, between the belief that p and either the belief that the evidence shows that not-p is more likely to be true than p or the belief that the requirement of total evidence should govern inductive belief formation. In addition to the absence of a certain kind of rational connection, Davidson requires the presence of a certain amount of rational organisation in the sub-systems for a partition to be genuinely in place. Thus Davidson writes

> What is called for is organised elements, within each of which there is a fair degree of consistency, and where one element can operate on another in the modality of non-rational causality (Davidson (1985), p. 301, for further discussion see pp. 301-305).

Dividing an agent into sub-systems by itself does not run counter to the interpretist approach. It is no part of this approach that all the objects of interpretation must exactly correspond to the intuitive boundaries of agents. The existence of intra-agent irrationalities suggests that it must recognise something smaller: sub-agencies of certain kinds. Once the division has been made, intra-agent irrationalities are no more puzzling than you believing that it is sunny and I believing it is not. Our feeling that we can make sense of self-deception and relative confidence in the attributions it involves is explained by the thought that we have something like this view in mind. There is, of course, a residual concern that the approach loses the unity of the person but I do not intend to press that point here.

The discussion of the Interpretists' position up to this point has just focused on how the attribution of any one of the sets of propositional attitudes characteristic of self-deception is compatible with the interpretist approach. It does not provide an explanation of the irrationality. However, their view of mental partitioning has consequences for the way they choose to explain irrationality. If the mental partition is merely to be understood in terms of the *absence* of a certain kind of rational connection and the presence of a certain amount of rational organisation in each sub-system, then it is pretty clear that the explanation of the irrationality will lie elsewhere.[3] Identi-

[3] Pears suggests that we should adopt the following negative criterion: an element is assigned exclusively to a sub-system if and only if it fails to interact rationally with any

fying mental partitions so understood is just a way of redescribing the very phenomenon we seek to explain (as Davidson acknowledges, see Davidson (1985), p. 147, contrary to what Heil seems to suppose, see Heil (1989), pp. 581-582). If Interpretists take the explanation of this type of irrationality seriously (as they should), they will inevitably be partitioners in the specified sense but they will be sceptical about appeals to partitions to explain the irrationality.

In particular, Interpretists are sceptical about the explanatory pretensions of richer notions of mental partitions appealing to consciousness. The Interpretist does not have to deny that if the agent had really focused on the evidence then it would have been impossible for him or her to believe that p. The issue is whether the description in terms of attentive consciousness is essential to the characterisation of what is going on or whether the Interpretist's characterisation cuts deeper. The Interpretist claims that the agent has the motivationally favoured belief *because* of the activity of the sub-system and that the agent's failure to be attentively conscious of the evidence against it is a consequence of this activity and the rational discontinuities which result.[4] The mental partition may coincide with the presence or absence of attentive consciousness but the explanatory weight is carried elsewhere (see Davidson (1982), p. 304; Davidson (1985), p. 147, fn. 10). Thus, Pears writes of Davidson's approach: 'It does not make any use of consciousness in drawing the line between main system and sub-system' (Pears (1984, 1986), p. 83; see also Davidson (1985), p. 147).

More precisely, the Interpretist explains the irrational combination of beliefs mentioned above by ascribing to a sub-system of the agent a distinctive rational agency which acts to produce the belief. The sub-system includes the desire that the agent believes that p and, if necessary, a means-end belief about how this may be brought about. For instance, the means-end belief

---

element in the main system (Pears (1984, 1986), pp. 97-98). This is clearly too strong. I very much doubt whether any belief fails to interact rationally with any other element of the main system. In which case, we lose the role that partitioning is meant to play in preserving Interpretism in the face of irrationality. Hence I have adopted a slightly looser conception of partitioning. The absence of *a* rational connection plus the presence of rational integration on the other side of the partition make it appropriate to assign a belief to the other side of a partition. This is more obviously in line with Davidson (1985), p. 147.

4 It is at this point that a difference will emerge between Davidson and Dennett's view of Interpretism. According to Davidson, beliefs and desires are causes. Therefore, Davidson will take this sentence to be true because the sub-systemic beliefs and desires combine to cause the attention of the subject to be focussed on other things than the evidence. By contrast, Dennett takes beliefs and desires to be abstracta and, hence, to have no causal role. In which case, the sentence will not be true in virtue of the efficacy of the beliefs and desires attributed to the sub-system. Instead, these beliefs and desires provide us with a way of making sense of the activity of the sub-system. The difference will not effect the conclusions that follow.

may be that, if the subject can be made to focus on the one slender piece of evidence for p, which is in fact outweighed by all the other evidence, then the subject will believe that p (Davidson (1985), pp. 145-148).[5]

Fido and his ilk have both the required rational discontinuities and sub-agency to be susceptible to the kind of explanation just identified. So Interpretists have no trouble with explaining Fido. The problem they have is in explaining the difference between Fido and the self-deceived. I have argued that this difference can only be captured by appeal to two things: first, the instability of the beliefs of the self-deceived when subject to the glare of attentive consciousness; second, the fact that the self-deceived are, in fact, not conscious of the presence of, or operation of, certain of their beliefs and desires. Fido displays neither of these features. The question is whether Interpretism is threatened by its inability to capture the difference between self-deception and Fido-cases.

There are many things that Interpretists can't explain, for example, the distinctive patterns of belief and desire when inebriated, very tired, or just plain careless. That does not count against the interpretive approach. Interpretists distinguish between personal and sub-personal psychology. Personal psychology is geared at the explanation of manifestations of agency in terms of beliefs and desires. Sub-personal psychology concerns how beliefs, desires, and other aspects of our mental life which explain our actions, are realised. It also concerns the consequences of characteristics of this realisation (see Dennett (1981b), pp. 57-65). If we drink a lot, our brains function less well, powers of reason break down and our behaviour becomes less and less amenable to explanation via the interpretive approach. Explanation of such behaviour lies outside the interpretive net and should focus, instead, on features of the brain and lower level cognitive processes (or so the story goes).

Obviously more detail is needed. However, we have enough to get in view the basic challenge that Fido-cases present. Unlike inebriation, tiredness and the like, self-deception appears to involve the further expression of agency rather than a departure from it. Identification of self-deception also seems to be something that is part of our explanation of agents at the personal level. It is not just inebriation. Self-deception presents a problem for the Interpretist precisely because it is something which rightfully seems it should be both understood and, in turn, play an explanatory role at the personal level.

Here's one rather cheeky response Interpretists might be inclined to make. They could concede that they do not manage to capture the difference between Fido-cases and the self-deceived. Nevertheless, they may claim that this differ-

---

[5] Although Pears and Davidson share the view that the activity of the sub-system is to be explained by appeal to rational agency, they disagree over whether the sub-system itself should be seen as a rational agent. Davidson says 'no' (Davidson (1982), pp. 303-304). Pears says 'yes' (Pears (1984), Ch. 5, Pears (1991), p. 396). I do not believe that a resolution of this difference touches on the points made in my paper.

ence is explanatorily inessential. It is a strength of the Interpretist position that it can explain both Fido-cases and the self-deceived in the same way without appeal to attentive consciousness. If it turns out to be explanatorily inessential, then maybe there is no harm in suggesting that self-deception should be reclassified as, partly, a sub-personal phenomenon.

The cheeky response won't work. The problem is that the presence or absence of attentive consciousness is not merely an add-on. It has explanatory ramifications. If the partition drawn in terms of rational discontinuities and organisation had coincided with that drawn in terms of attentive consciousness, we would have had grounds for supposing that attentive consciousness was either the result of the application of reason or the vehicle of the application of reason. In that context, it would have been perfectly appropriate to characterise beliefs and desires by their role in a normative interpretive scheme. Once we see that attentive consciousness is independent then a number of questions become relevant, for instance: Are some beliefs and desires more likely to involve attentive consciousness? Are certain irrational belief-desire combinations more unstable in the face of attentive consciousness than others? Are these factors constant across individuals? and so on. This suggests that appeal to a host of supplementary ceteris paribus laws will be required to engage in the proper attribution of beliefs and desires. In which case, beliefs and desires are not to be understood just in terms of their role in a normative interpretive scheme. Indeed, once one allows appeal to facts about belief and desire outside the normative framework of interpretation, it is hard to know where to stop. Rationality and the good will lose their constitutive status in interpretation in favour of ceteris paribus laws concerning beliefs, desires and the like, which serve to aid the ascription of mental states. Of course, some of these ceteris paribus laws will include the transitions which would happen in the minds of rational subjects and lovers of the good, applying ceteris paribus to the rest of us, but their status would be much reduced.

I am not the first to challenge Interpretism as it has been understood here. For instance, Alvin Goldman has provided cases in which we would be inclined to avoid trying to maximise the rationality of those we seek to interpret (Goldman (1989), pp. 10-13). These cases involve the paradox of the preface and the mistakes in probabilistic reasoning identified by Amos Tversky and Daniel Kahneman (Tversky and Kahneman (1982), pp. 91-96; Tversky and Kahneman (1983)). The difference is that the challenge I pose involves something which, by Interpretists' own lights, Interpretism should be able to explain. Errors of reasoning of the kind Goldman mentions can be explained away by subpersonal factors. Some may even be explained without loss by the kind of explanation favoured by Interpretists which, I have argued, fails to capture what is going on in the case of self-deception.

Let me make clear that the challenge I am posing the Interpretist does not imply a rejection of interpretation more broadly conceived. If interpretation is understood as the attribution of content-bearing states, like beliefs and desires, to subjects governed by our psychological knowledge, then nothing that I have written undermines this practice (and so not to *Interpretism Psychologised*, as one might put it). However, this is not the doctrine defended by Davidson and Dennett amongst others.

One response that might appear open to the Interpretist is to adopt a stratified theory. Suppose we find that a reductive account of attentive consciousness is available. To fix ideas, let us consider the following toy theory of attentive consciousness: S is attentively conscious of X if and only if S judges that S is in a mental state with X as its object. If it turned out that our attribution of these higher-order judgements was governed by the Interpretist's framework, then the facts about attentive consciousness would be available to the Interpretist after all. Even if it turned out that there were ceteris paribus laws governing how we should interpret agents in the light of attributions to them of attentive consciousness of some things and not others, the interpretive framework would still prove fundamental. It would provide the first tranche of interpretive information on which to work.

This would already be a significant retreat from Interpretism. But, more important, it is unlikely that it will work. There is no reason to suppose that the attributions of higher order judgements necessary to get the attribution of attentive consciousness right will be mandated by a normative scheme of interpretation. Initially, it might seem plausible that they would be mandated. There is little doubt that, if I consciously attend to the fact that p, then it is rational for me to judge that I am in a mental state with the fact that p as its object. But this does not provide independent interpretive grounds for attributing to me the higher-order judgement. It requires the prior identification of states involving attentive consciousness in order to settle what it would be rational for me to judge. There may be no reason to ascribe to me the higher order judgement other than the fact that I am attentively conscious of the fact that p. Indeed, a common charge against higher-order thought theories of consciousness is that they propose to ascribe to subjects mental states which there is otherwise no reason to ascribe to them (see Rosenthal (1986); Davies and Humphreys (1993), pp. 23-27; Dretske (1995), pp. 110-112). This promises to be just as much a problem for our toy theory of attentive consciousness.

If this problem faces our toy theory, it will be much worse for those which don't involve the attribution of mental states which are supposed to be characteristic elements in a normative interpretive scheme. For instance, suppose the claim was just that a subject should be *disposed* to make the higher-order judgement, rather than actually make it, or that states of attentive con-

sciousness should just have a certain kind of cerebral celebrity through successfully dominating the resources of memory, utterance, reflection and the like (Dennett (1991); Dennett (1993), p. 929). Then it is really quite unclear how the interpretive scheme would enable us to attribute states of attentive consciousness. My point is not that there cannot be a reductive theory of attentive consciousness. The point I am making is quite neutral on that. My point is just that it is unlikely that we would be able to appeal to such a theory to explain how the interpretive approach will attribute states of consciousness.

There might appear to be an obvious adjustment to the interpretist approach which would deal with the problem raised by Fido-cases. Part of the support for Fido's belief that she still loves him seems to derive from the importance he places upon faithfulness to the relationship. I have been working on the assumption that the normative requirements that govern belief formation concern evidence. Perhaps there are also practical requirements upon belief formation deriving from our deepest commitments, the faithfulness that Fido feels to a relationship being one such example. We might then think of the *overall requirement on belief formation* as this

One gives credence to the hypothesis that one ought to believe.

Often the belief recommended by this requirement will be the same as that recommended by the requirement of total evidence. Sometimes their recommendations come apart due to the influence of practical considerations. This happens in the case of Fido.

If this suggestion were correct, then we should make a corresponding adjustment to the Davidson-Pears understanding of the notion of a mental partition. A mental partition is present only if one forms a belief against the overall requirement of belief formation. In which case, there is no partition in Fido's mind and, hence, it is no surprise that his allegedly conflicting beliefs can both involve attentive consciousness. His belief that she still loves him is in line with the overall requirement and not in conflict with the belief about the evidence. If this were right, Interpretism could appeal to attentive consciousness to characterise self-deception and yet cash this out in terms of their favoured notion of partitioning. There would have to be no independent appeal to attentive consciousness.

Unfortunately, even if the point about the requirement on belief formation is correct, the corresponding adjustment to the notion of a mental partition doesn't save the interpretist approach. Suppose there is a weak husband who loves his wife very much but feels a strong attraction to a woman he has recently met. In the way these things tend to happen, he believes that the evidence shows that it is more likely that she is attracted to him than that she is not. He believes that it is right to remain faithful to his wife but he is

weak. He believes that, if he believes that the woman is attracted to him and the opportunity presents itself, he will be unfaithful. On the other hand, if he believed that she found him unattractive, he would be out of danger. It strikes him as much more important for him to remain faithful to his wife than to believe the truth in these circumstances. It would appear that he ought to believe that the woman doesn't find him attractive even though the evidence is strongly to the contrary. However, he finds that he can't form this belief. In his view, there is not enough evidence.

In this case, it seems just as legitimate (if not more so) to suppose that the overall requirement of belief formation departs from the requirement of total evidence. If failures to follow the overall requirement corresponded to failures of attentive consciousness, then the unfaithful husband ought to fail to be attentively conscious of *either* the fact that the overall requirement of belief formation requires the belief that she does not find him attractive *or* the fact that she finds him attractive. However, it seems all too plausible that the weak husband is attentively conscious of both. So the adjustment to the requirement on belief formation doesn't seem to get round the problem.

A natural thought to have is that the weak husband is ambivalent. Maybe he thinks it important not to betray his wife but maybe he also wants the other woman to find him attractive. That's why he fails to form the belief. I have little doubt that this could be part of the explanation but it does not touch the point I just made unless his desire that she find him attractive has sufficient *normative* force to tip the balance in favour of the requirement of total evidence. This does not seem to be so.

There are quite a few circumstances in which we may reflectively judge that it is more important to have a belief that p in the service of something we value than the belief that not-p mandated by the evidence. It is puzzling, then, that cases of this sort are not more prevalent bearing in mind that they do happen sometimes. It seems that attentive consciousness has the tendency to privilege the requirement of total evidence out of proportion to its normative weight in belief formation as revealed by our reflective judgements on the matter. Indeed, contrary-to-evidence belief formation is likely to strike us as peculiar partly because we are naturally inclined to think about cases involving attentive consciousness. We think about whether we could form a belief that p in the face of evidence of such and such a character. These are the cases which are least likely to display contrary-to-evidence belief formation (for much more discussion on this and related issues see Noordhof (2001)).

If attentive consciousness has a predilection for the requirement of total evidence in spite of the fact that we may reflectively judge that the weight should be placed elsewhere, then this poses a further difficulty for Interpretism. Belief and desire are supposed to be ascribed within a normative interpretive framework. Attentive consciousness's favouring of the requirement of

total evidence results in a departure from what, overall, one ought to believe on certain occasions. Successful interpreters should appeal to what actually governs conscious and non-conscious reasoning, and not what ought to govern it, in attributing beliefs and desires. The resulting interpretation will be perfectly intelligible to you or me given that we reason in the same way but it will not have the character which Interpretists insist that it must have.

The connection between attentive consciousness and the requirement of total evidence provides further reason to be sceptical about the idea that self-deception should be reclassified as a partly sub-personal matter. It is not just that appeal to attentive consciousness proves explanatorily relevant in the explanation of action. It is rather that its explanatory relevance is tied to the influence of principles of belief and desire formation. These, in turn, relate to issues of agency and control.

Fido and his kin reveal that the requirement of total evidence does not always prevail in attentive consciousness. However, its defeat seems to require something which is powerfully rooted in an agent's current self-conception. This places some cases of self-deception in a rather different light. Self-deception may occur as a means of getting round attentive consciousness's tendency to favour the requirement of total evidence. Often such deception will be at the service of unworthy ends but sometimes it might be to the good. The weak husband, for instance, might be better off self-deceivedly believing that she does not find him attractive rather than facing the awful truth. In such cases, self-deception is not plausibly seen as a loss of self control. Self-control does not require that we always follow the warped agenda of our conscious mind at the expense of everything else we value. Self-control can involve an agent having sufficient resilience to pursue certain ends—for example, remaining faithful to one's wife—aided by a well-placed bit of self-deception. It would, of course, be better if this were not necessary. However, given the agent's dispositions, this might be the best way to retain control.

It might be thought misguided to appeal to attentive consciousness to explain some mental phenomenon when it is so clearly in need of further explanation itself. I do not share this rather strict view. I have given a preliminary specification of the phenomenon I have in mind by the term 'attentive consciousness'. The question, then, is how to proceed from there. Part of any theory of consciousness will attempt to elucidate it by drawing upon its relation to other things. This paper has just highlighted the fact that attentive consciousness is involved in the phenomenon of self-deception; that attentive consciousness favours certain requirements on belief formation at the expense of others; and that this may explain why self-deception occurs. These connections illuminate the character of attentive consciousness and should be part of any theory of its nature. It is perfectly appropriate to recog-

nise this even if attentive consciousness eventually receives some further more fundamental explanation.[6]

References

Robert Merrihew Adams (1995), 'Moral Faith', *Journal of Philosophy*, 92, pp. 75-95.

Robert Audi (1989), 'Self-Deception and Practical Reasoning', *Canadian Journal of Philosophy*, 19, pp. 246-266, and in his (1993), *Action, Intention and Reason* (Ithaca and London, Cornell University Press), pp. 209-230 [page references in text to latter].

Annette Barnes (1997), *Seeing Through Self-Deception* (Cambridge, Cambridge University Press).

J. V. Canfield and D. F. Gustavson (1962), 'Self Deception', *Analysis*, 23, pp. 32-36.

Rudolf Carnap (1950), *The Logical Foundations of Probability* (London, Routledge and Kegan Paul, Ltd.).

William Child (1994), *Causality, Interpretation and the Mind* (Oxford, Oxford University Press).

Donald Davidson (1970), 'Mental Events', in his *Essays on Action and Events* (Oxford, Oxford University Press) pp. 207-225.

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

Donald Davidson (1983), 'A Coherence Theory of Truth and Knowledge', Ernest LePore (ed.), *Truth and Interpretation* (Oxford, Basil Blackwell), pp. 307-319.

Donald Davidson (1985), 'Deception and Division', in E. LePore and B. McLaughlin (eds.), *Actions and Events* (Oxford, Basil Blackwell), pp. 138-148.

Martin Davies and Glyn W. Humphreys (1993), 'Introduction', in their (1993, eds.), *Consciousness* (Oxford, Basil Blackwell), pp. 1-39.

Daniel C. Dennett (1981a), 'True Believers: The Intentional Strategy and Why It Works', A. F. Heath (ed.), *Scientific Explanation* (Oxford, Oxford University Press) reprinted in his (1987), *The Intentional Stance* (Cambridge, Massachusetts, The MIT Press), pp. 13-35.

---

[6] A very distant ancestor of this paper was read at the Moral Sciences Club Cambridge, the Sheffield Philosophy Department Seminar, and U.C.L. Philosophy Society. It goes without saying that the audiences enabled me to improve the paper considerably for which I am very grateful. Special thanks are due to Michael Clark, Jim Hopkins, Jen Hornsby, Bob Kirk, Mike Martin, Greg McCulloch, Hugh Mellor, Tom Pink and two anonymous referees for PPR. I would also like to thank Clare Hall, Cambridge, for providing a Research Fellowship during which I began work on this paper, and more recently, the Mind Association for a Research Fellowship and the AHRB Matching Research Leave Scheme for support.

Daniel C. Dennett (1981b), 'Three Kinds of Intentional Psychology', R. Healy (ed.), *Reduction, Time and Reality* (Cambridge, Cambridge University Press), reprinted in his (1987), *The Intentional Stance* (Cambridge, Massachusetts, The MIT Press), pp. 43-68.

Daniel C. Dennett (1991), *Consciousness Explained* (London, Allen Lane).

Daniel C. Dennett (1993), 'The Message is: There is no *Me*dium', *Philosophy and Phenomenological Research*, 53, no. 4, pp. 919-931.

Fred Dretske (1995), *Naturalizing the Mind* (Cambridge, Massachusetts, The MIT Press).

Herbert Fingarette (1998), 'Self-Deception Needs No Explaining', *The Philosophical Quarterly*, 48, no. 192, pp. 289-301.

Alvin Goldman (1989), 'Interpretation Psychologised', Mind and Language, 4, pp. 161-185, and in his (1992), *Liaisons* (Cambridge, Massachusetts, The MIT Press), pp. 9-33 [page references in text to latter].

John Heil (1989), 'Minds Divided', *Mind*, 98, no. 392, pp. 571-583.

Carl G. Hempel (1965), *Aspects of Scientific Explanation* (New York, The Free Press).

Richard Holton (1994), 'Deciding to Trust, Coming to Believe', *Australasian Journal of Philosophy*, 72, no. 1, pp. 63-76.

Mark Johnston (1988), 'Self-Deception and the Nature of Mind', in B. McLaughlin and Amélie Oksenberg Rorty (eds.), *Perspectives on Self-Deception* (Berkeley, University of California Press), pp. 63-91.

Sören Kierkegaard (1846), *Concluding Unscientific Postscript* (New Jersey, Princeton University Press).

Christine Korsgaard (1986), 'Skepticism about Practical Reason', *Journal of Philosophy,* 83, no. 1, pp. 5-25.

Ariela Lazar (1999), 'Deceiving Oneself or Self-Deceived? On the Formation of Beliefs "Under the Influence"', *Mind*, 108, no. 430, pp. 265-290.

Alfred R. Mele (1987a), *Irrationality* (Oxford, Oxford University Press).

Alfred R. Mele (1987b), 'Recent Work on Self-Deception', *American Philosophical Quarterly*, 24, no. 1, pp. 1-16.

Alfred R. Mele (1997), 'Real self-deception', *Behavioral and Brain Sciences*, 20, pp. 91-136.

Adam Morton (1988), 'Partisanship', in Brian McLaughlin and Amélie Oksenberg Rorty (eds.), *Perspectives on Self-Deception* (Berkeley, University of California Press), pp. 170-182.

Paul Noordhof (1999) 'Review of Annette Barnes, *Seeing Through Self-Deception*', *Philosophical Books*, 40, no. 3, pp. 180-184.

Paul Noordhof (2001), 'Believe What You Want', *Proceedings of the Aristotelian Society*, 101.

David Pears (1984,1986), *Motivated Irrationality* (Oxford, Oxford University Press).

Terence Penelhum (1964), 'Pleasure and Falsity', *American Philosophical Quarterly*, 1, no. 2, pp. 81-91.

David M. Rosenthal (1986), 'Two Concepts of Consciousness', *Philosophical Studies*, 49.

Dion Scott-Kakures (1996), 'Self-Deception and Internal Irrationality', *Philosophy and Phenomenological Research*, 56, no. 1, pp. 31-56.

Michael Smith (1994), *The Moral Problem* (Oxford, Basil Blackwell).

Bela Szabados (1973), 'Wishful Thinking and Self-Deception', *Analysis*, pp. 201-205.

William Talbott (1995), 'Intentional Self-Deception in a Single Coherent Self', *Philosophy and Phenomenological Research*, 55, no. 1, pp. 27-74.

Amos Tversky and Daniel Kahneman (1982), 'Judgements of and by Representativeness', in Daniel Kahneman, Paul Slovic and Amos Tversky (eds., 1982), *Judgement under uncertainty: Heuristics and biases* (Cambridge, Cambridge University Press), pp. 84-98.

Amos Tversky and Daniel Kahneman (1983), 'Extensional versus Intuitive Reasoning: The Conjunctional Fallacy in Probability Judgement', *Psychological Review*, 90, pp. 293-315.

Miguel de Unamuno (1921), *Tragic Sense of Life* (New York, Dover Publications, Inc).