doesn't refer to actual BIV, but to something else (such as BIV in the image, stimuli, program features, or nothing at all). Let us call this other referent 'VIB', whatever it is). Thus, the objection goes on, even his words do not refer to actual BIV. But (still assuming we are BIV) the objection is beside the point. For we must realize that when the sceptic thinks or says 'we might be BIV', he means 'we might be VIB' and that is precisely what he intends to say. He is not trying to say anything of actual BIV.

The opponent might be satisfied with this answer, and think that is concedes his own point. But there is an appendix to this, which has to do with the topic of the preceding section: even the opponent himself (assumed things are as he says) by 'BIV' actually refers to VIB. Hence, when he claims

(*) By 'BIV' the sceptic does not refer to actual BIV

he is actually claiming that the sceptic does not refer to actual VIB, which is false according to Putnam's theory of reference. In fact, (*) is false according to any theory of reference (although different theories of reference will give different readings of it). For example, according to a description theory of reference, even if uttered by BIV, (*) means simply that by 'BIV' the sceptic does not refer to actual BIV. And this is false by the description theory.

According to any theory of reference, (*) is more than false, it is self-contradictory. For it says that a certain term in the sceptic's language ('BIV') does not refer to a certain object; but it says that in the same language (whether it be regular English or vat English)[25] and refers to that object by the same term. Thus, as we saw in section 6, in effect it says that 'BIV' does not refer to the object to which it refers.

In conclusion, any attempt to show that the BIV hypothesis is incoherent fails. Philosophers may go on forever discussing what is reference, what is mind, and whether the BIV would have one or the other. But scientists need not wait their response to decide whether BIV are possible.

Via Curiel 41
I-47023 Cesena (FO), Italy

25. For we have just noticed that the sceptic and his opponent are in the same linguistic boat: they are both either BIV or regular persons, they both either succeed or fail to refer to external things.

ALEXANDER BIRD

Rationality and the Structure of Self-Deception

Abstract. This paper concerns techniques of psychological explanation, in connexion especially with irrationality. My prime example is self-deception. In order to explain this phenomenon, Freud, and philosophers since, in particular, Davidson, have adopted a model of the mind as compartmentalized in such a way as the parts of the mind act as semi-autonomous minds themselves, thereby allowing an explanation of self-deception modelled on ordinary, public other-deception. I argue that such a model of the mind is highly problematic, and that self-deception can be accounted for without it by observing that the essence of self-deception is the presence of a motive for possessing a false or irrational belief. The primary sources for criticism are thus Freud and Davidson. Motivation for such criticism derives from Sartre's attacks on Freud.

1.

The focus of this paper is the explanation of self-deception. I shall consider accounts of its structure and of its relation to rationality and irrationality. In particular I aim to criticize accounts of self-deception which appeal to an other-deception model for self-deception, notably the Freudian picture of a divided mind. In the context of Freudian psychoanalysis, self-deception is seen as a tension within the mind of the patient and a matter of pathology. Self-deception requires an internal repression which is a drain on a person's energies and resources. The repressed person's existence is not authentic. He is true neither to himself nor to others. By releasing him from this, the energy otherwise expended in maintaining this internal struggle can be used to broaden the subject.[1]

1. See S. Freud, *Collected Papers*, trans. J. Riviere, London: Hogarth Press, 1953, vol.iv, pp. 89–90, and I. Dilman, *Freud and the Mind*, Oxford: Oxford University Press, 1984, p. 95.

Such a view sees self-deception as destructive, a perversion. Perhaps it is a feature of late nineteenth and twentieth century liberalism that Freud should see this in terms of pathology while two centuries earlier Bishop Butler was inclined to give a moral construction, condemning self-deception as an internal hypocrisy which enables the wicked man not only to behave in an evil fashion, but also to hide from himself the awfulness of his deeds.

Yet reflection upon instances of self-deception should reveal that not all involve inner conflict or are worthy of moral contempt. We may not find abhorrent or destructive the faith of a caring parent in the innocence of its child when faced suddenly with evidence of its previously hidden viciousness. A lay observer may feel indeed that this is the right reaction for a good parent. A sympathetic analyst might regard this as a positive feature, allowing a slow arrival at the truth, where a sudden realization would be intensely damaging.

This is what Sartre had in mind when discussing *mauvaise foi*—bad faith.

> To be sure, the one who practices bad faith is hiding a displeasing truth or presenting a truth as a pleasing untruth. Bad faith then has in appearance the structure of falsehood. Only what changes everything is the fact that in bad faith it is from myself that I am hiding the truth. Thus the duality of the deceiver and the deceived does not exist here. But on the contrary, implies in essence the unity of a single consciousness.[2]

This suggests the answer to what must be our first question: why should Freud and Butler have seen self-deception as a pathological division of the mind or an immoral hypocrisy, when our examples seem to suggest that there are cases where the opposite is true? The Freudian or what I shall call *divisionist* picture employs a super-ego or censor whose role is to prevent the knowledge of the id reaching the ego. It shares this with Butler's view, that self-deception is depicted as an internalized version of other-deception. When *A* deceives *B*, *A* knows that some proposition *p* is false and deliberately brings is about that *B* believes that *p*. The other-deception model of self-deception has it that *A* is one and the same as *B*. *A* really knows that *p* is false but makes himself believe that *p* is true. Self-deception is just like ordinary other-deception, excect that it is a special case where the deceiver is identical with the deceived.

Yet this would imply that *A* both believes *p* to be true and believes *p* to be false. This underlying fact of the other-deception model explains why Butler takes self-deception to be hypocritical, for the belief that *p* will lead the self-deceiver to behave in a manner contrary to his other belief that not-*p*.

2. J.-P. Sartre, *Being and Nothingness: An Essay in Phenomenological Ontology*, New York: Philosophical Library, 1956, Part I, chap. 2.

However the ascription to one believer both of a pair of contradictory beliefs is problematic. There are certain circumstances which might make such an ascription plausible, e.g., where the beliefs are separated in time, but none of these seem to suit all cases of self-deception (see below p. 25).

Furthermore, it is part of (successful) other-deception that the deceiver knows he is deceiving but that the deceived does not know he is being deceived. But if the deceiver and the deceived are identical, then it must be that he both knows he is deceiving himself and does not know he is deceiving himself. Successful self-deception, seen as a special case of other-deception, generates a problem of accounting for contradictory beliefs as well as a paradox, that the self-deceiver both knows that he is deceiving and does not know that he is deceiving.

Of course, one may believe *p* at one time and not-*p* at another time, and these states may even alternate in a being with a genuinely split personality. However it is questionable whether it is possible for a person to be in both belief states at the same time, because the criteria for believing *p* and believing not-*p* will generally be mutually exclusive. If I wish to quench my thirst then I will drink the glass. If I believe it will not quench my thirst then, if I have no other reason for drinking it, I will not drink the glass. The behavioral criteria for the two beliefs, drinking and desisting from drinking, are clearly incompatible.

It is to avoid these problems of the simple other-deception model that Freud introduces the divisionist element. Now it is not the case that the self-deceiver literally deceives himself just as he might deceive another; rather one part of the self-deceiver, the id (or id plus super-ego) deceives another, the ego, where the deception practiced by one part of the mind on the other is precisely analogous to genuine other-deception.

It is my view that this increase in the complexity of the other-deception model for self-deception serves only to mask its difficulties, and not to exclude them. At the same time it introduces further confusions. In the following I shall enlarge upon this claim, in particular in relation to Davidson's account of self-deception. I shall conclude by showing that an account of self-deception which does not invoke the other-deception model and therefore does not invoke a picture of the mind as divided can satisfactorily characterize the phenomenon. One can satisfy Sartre's demand that one portray the self-deceiver as possessing a single consciousness and a unity of purpose. Essentially one need only point to the fact that the false belief maintained by the self-deceiver is a motivated belief.

2.

The first problem with the divisionist account is that it is in danger of introducing the homunculus fallacy, that the mind (or brain) is like a person within a person, the possessor of that mind. Here there is within a person not just one homunculus but several. We have talked of the id and the ego having beliefs and the one being deceived by the other as if they were human beings. But they are not human beings, being parts of the mind of one; in which case what sense can be made of the ascription of human attributes, viz possession of beliefs, to them? For the criteria for the correct ascription of these attributes such as belief make reference to the behaviour of the possessor, but such behaviour can be exhibited only by the complete human being. It makes no sense to suggest that a mind (or brain, for that matter) or part of one can exhibit the behaviour characteristic of the belief that such and such.

If we do wish to talk of a mind or brain having beliefs it must then be parasitic on the ascription of belief to the possessor of that mind or brain. That is, to say that *X*'s mind (brain, ego, id, etc.) believes that *p* is to say that *X* believes that *p*, i.e., that *X* satisfies the relevant behavioural criteria. This entails that if *X*'s ego believes that *p* and at the same time his id believes that not-*p*, then at that time, *X* both believes that *p* and believes that not-*p*. If to say that *X*'s ego believes that *p* is to say that *X*'s behaviour which is criterial of that belief is caused by the ego, while, similarly, *X*'s id causes behaviour appropriate to the belief that not-*p*, then the question arises how can both the ego and the id cause the behaviour appropriate to each of their contradictory beliefs? It would appear then, that insofar as my remarks above do point to problems with the simple other-deception model, then those same problems are not eliminated but are merely obscured by the more sophisticated version.

Of course the conflict of behavioural criteria will not arise permanently, but only in certain relevant circumstances. If there is no opportunity of drinking the glass of water, then there will be no opportunity to display one's beliefs about its thirst-quenching properties. In such circumstances there will not be any conflict between contradictory beliefs, since the circumstances in which the contradictory forms of behaviour would occur do not here arise. However, if one believes that the correct ascription of beliefs is connected with behaviour in certain counterfactual situations, then the problem remains. What a person's present belief concerning thirst-quenching by the water is, will be shown by the answer to the question: how would he behave if he were thirsty and could drink the water?—to which question the answer will be appropriate to only one of the beliefs and not the other. The counterfactual dependence of behaviour on beliefs and desires is more complex than this, but I would question whether there is any description of behaviour in counterfactual situations which is unambiguously indicative of holding contradictory beliefs. For instance, one might suggest that in some counterfactual circumstances *X*'s behaviour indicates the belief that *p*, and in others the belief that not-*p*. It is not clear that the right description is that the person holds contradictory beliefs, rather than that the person's beliefs are counterfactually dependent on the circumstances he finds himself in, or even that there is no definitive fact of the matter as to what the person does believe.

Even if one is to insist that some counterfactual descriptions are most appropriately regarded as the simultaneous possession of contradictory beliefs, such descriptions would not be those of self-deception. For the failure of a belief to manifest itself in a relevant counterfactual situation would need to be explained in terms of some form of dormancy—in such circumstances the belief is causally inactive. But the point of the other-deception model of self-deception is that both of the contradictory beliefs are active. The deceiver, qua deceiver, believes that *p* and brings it about that the deceived believes that not-*p*. Neither the deceiver or deceived has a dormant belief.

Of course, if one were to relax or deny the counterfactual relation between belief and behaviour, it would become possible to ascribe almost any set of beliefs to an individual. Certainly one extreme, Cartesian dualism, can accommodate contradictory beliefs nicely, and therefore the other-deception model for self-deception and a compartmentalized mind. But at the same time, the further one pursues the Cartesian road, the more difficult do one's conceptual problems of psychological explanation become. And it is for this reason that several of the various contenders in the philosophy of mind—logical behaviourism, the Wittgensteinian version of it, and the functionalist version, emphasize the sort of connexion between belief and behaviour I have adopted here, though I shall not argue this well-debated issue further.

In any case, given that the whole person can display at one time the criteria appropriate to a belief but not its negation, what grounds can there be to

ascribe the belief of the negation? The grounds are simply that the subject being rational (and his or her behaviour in other circumstances will show this) must believe such and such, since any contrary belief will show the subject to be irrational. Thus some part will know the truth, while some other believes something different. The id takes in the facts and comes to the correct conclusion. Roughly speaking, for its own purposes, the id selectively presents facts (or even falsehoods) to the ego which then comes to an incorrect conclusion, albeit rationally. It is this false conclusion which displays itself in behaviour. Clearly, maintaining the rationality of both ego and id depends on the fact that the ego has the fact presented to it selectively. But this simply does not fit the phenomenon of self-deception. For in a case of genuine self-deception (as opposed, e.g., to merely forgetting pertinent evidence) the self-deceiver can be made to acknowledge all the evidence while at the same time denying the obvious conclusion. In which case there is no room to claim that one part of the mind (the ego) denies the conclusion, while another (the id) alone acknowledges the evidence. If the evidence is represented by the proposition q, and the rational conclusion by p, then the self-deception can be represented by the single belief, the belief that (q & not-p), and when this single belief is expressed it must thereby be the belief of a single part of the mind. If the self-deceiver shows his or her self-deception by uttering: "I know that q, but I believe that not-p," then the question arises, which portion of the divided mind is causally responsible for this speech act. Whichever it is must be accorded the irrational belief that (q & not-p). In which case one part of the mind at least is irrational. But then there is no need to introduce separate parts of the mind, since this was done only in order to maintain rationality.

3.

Davidson is the most noted among contemporary philosophers to accept the Freudian or divisionist picture. He maintains that in self-deception not only does the subject hold incompatible beliefs, but also that the one belief causes its contrary. He writes:

> An agent A is self-deceived with respect to the proposition p under the following condition. A has evidence on the basis of which he believes that p is more apt to be true than its negation; the thought that p, or the thought that he ought rationally to believe p, motivates A to act in such a way as to cause himself to believe the negation of p. The action involved may be no more than an intentional directing of attention away from the evidence in favour of p; or it may involve the active search for evidence against p. All that self-deception demands of the action is that the motive originates in a belief that p is true (or recognition that the evidence makes it more likely to be true than not), and that the action be done with the intention of producing a belief in the negation of p. Finally, and it is especially this that makes self-deception a problem, the state that motivates self-deception and the state it produces coexist.[3]

Given what I have just remarked, it does not seem possible that two incompatible beliefs should interact in this way. On the one hand it is possible for my belief that p to cause me to act to bring it about that I believe that not-p. For instance I may have to undertake some dangerous act. I know that if I am aware of this at the time of performing the act, my resulting fear will increase the chances of failure. Thus knowing that the act is dangerous I may first act such as to bring it about that I think it is quite safe. But it seems that part of what success in this project involves is precisely that I finally no longer have the belief I started with. If I genuinely succeed in thinking the act safe, then I must have succeeded in forgetting that it is dangerous. This is in fact Pears' analysis of self-deception.[4] Deceiving one's future self may well be called self-deception. In favour of this view is that it does not entail any compartmentalization of the mind. However, deceiving one's future self is clearly different from many other cases of self-deception, where, on the divisionist account, contrary beliefs are held simultaneously. For (as Davidson recognizes) what seems to be central to such cases of self-deception is that the false belief is held in the face of contrary evidence; and this does not seem to be an element of deceiving one's future self, where, presumably, the past self will attempt to shield the future self from the evidence.

On the other hand, if the true belief is still clearly present (and not forgotten), then it cannot be that I have come genuinely to believe its negation also. For instance, I may, when walking a tightrope, persuade myself not to look down. Presumably this is in order to avoid coming 'face to face' with the fact that I am precariously high off the ground. One might call this a case of a belief causing me to act in such a way as to avoid direct contact with the fact that makes it true. But this simply amounts to my avoiding looking to see how precarious I am, and not to my coming to the belief that I am not precariously situated.

It would seem, so far as the interaction between the belief that p and the

3. D. Davidson, "Deception and Division," in J. Elster (ed.), *The Multiple Self*, Cambridge: Cambridge University Press, 1986, pp. 88–89.
4. See D. Pears, "Motivated Irrationality," in *Philosophical Essays on Freud*, ed. Wollheim and Hopkins, Cambridge: Cambridge University Press, 1982, pp. 264–288, and *Motivated Irrationality*, Oxford: Oxford University Press, 1984.

belief that not-p are concerned, that (i) if it is correct to say that my belief that p caused me to acquire the belief that not-p, then, in so acquiring this belief I must have relinquished the original belief that p; (ii) if, on the other hand, there is good ground for saying that I still possess the belief that p (i.e., this belief continues to influence my behaviour, e.g., by preventing the tightrope walker from looking downwards, or stepping off the rope, etc.), then it cannot correctly be said that I also possess that belief that not-p.

Davidson gives this as his example:

> Carlos has good reason to believe he will not pass the test for a driving licence. He has failed the test twice before and his instructor has said discouraging things. On the other hand he knows the examiner personally, and has faith in his own charm. He is aware that the totality of the evidence points to failure. Like the rest of us he normally reasons with the requirement of total evidence. But the thought of failing the test once again is painful to Carlos (in fact the thought of failing anything is particularly galling to Carlos). So he has a perfectly natural motive for making it the case that he is a person who will (probably) pass the test. His practical reasoning is straightforward. Other things being equal, it is better to avoid pain; believing he will fail the test is painful; therefore (other things being equal) it is better to avoid believing he will fail the test. Since it is a condition of his problem that he take the test, this means it would be better to believe he will pass. He does things to promote this belief, perhaps obtaining new evidence in favour of believing he will pass. It may simply be a matter of pushing the negative evidence into the background or accentuating the positive.[5]

We can pick out two important elements in Davidson's account of the case:

(i) Carlos deceives himself into thinking that he will pass, because it is painful to think that he will fail;
(ii) Carlos is aware that the evidence favours the belief that he will fail.

That Carlos believes that he will pass, while (ii) is also true, means that Carlos is guilty of weakness of the warrant, holding a belief unsupported by the evidence. In fact the truth of (ii) means, in Davidson's view, that Carlos must also believe he will fail. In fact, Davidson says that he agrees that self-deception requires the simultaneous entertainment of incompatible beliefs. And so

(iii) Carlos believes he will fail.

Davidson holds that Carlos' belief that he will fail must coexist with the belief which is the outcome of the successful self-deception, (i), viz. that he will pass, since the former is Carlos' motivation for the latter. He writes: "[...] core cases of self-deception demand that Carlos remain aware that his evidence favours the belief that he will fail, for it is awareness of this fact that motivates his efforts to rid himself on the fear that he will fail."[6]

Yet, even laying intrinsic problems of the postulation of incompatible beliefs aside, this account cannot succeed. For Carlos' motivation in this self-deception is that (a) he believes he will fail plus (b) the fact that the belief (a) is painful. Davidson maintains that it is essential to self-deception that the belief that he will fail remains. But in this case, even if the self-deception succeeds in giving Carlos the belief that he will pass, the strategy of avoiding pain has failed, since he also still has the painful belief that he will fail. But since this belief can coexist, according to Davidson, with the belief that he will pass, acting so as to acquire the latter belief will do Carlos no good. It can do Carlos good only if it cannot coexist with the belief that he will fail. And so Davidson cannot both maintain that the beliefs may coexist and describe Carlos' self-deception as the strategy of avoiding the pain of one belief by acquiring the other.

It would seem that in order to make sense of the case, one might say

(I) Carlos' motivation is (as in Davidson's account) (a) the belief that he will fail, plus (b) the fact that this belief is painful. In which case his aim must be to rid himself of this belief. As Davidson admits, this must amount to Carlos' acquiring the belief that he will pass (i.e., deceiving himself). But to say this is to accept that incompatible beliefs may not coexist. Either way the success of the strategy must lie in Carlos' no longer believing he will fail.

Alternatively,

(II) Carlos' motivation is the pleasure that the belief that he will pass will give him. If he succeeds in acquiring this belief he will be deceiving himself. But it is no part of understanding this process that Carlos at any point believes that he will fail. (And indeed if it is right that not believing he will fail is a criterion of believing he will pass, then it cannot be that he believes that he will fail.)

Thus the coexistence of incompatible beliefs is not simply a problem in itself for Davidson. Above I questioned whether such a coexistence is possible; but even if we allow that it is, Davidson's acceptance of it as an element in his account of self-deception serves to vitiate that account. Self deception,

5. Davidson, D., "Deception and Division," art. cit., pp. 89–90.

6. Ibid p. 90.

it is agreed, involves the acquisition of a false or unwarranted belief. If coexistence of incompatible beliefs is allowed then the motivation of self-deception cannot be the desire to avoid the pain of some belief held presently; for by coexistence the acquisition of further beliefs cannot rid me of or eliminate previously held beliefs. On the other hand if the motivation for self-deception is the pleasure brought by the false belief, then the existence of any true contrary belief simply need not enter the story.

For Davidson self-deception amounts to weakness of the warrant, where additionally the warranted belief motivates the unwarranted belief. Such weakness of the warrant in self-deception he takes to be parallel to weakness of the will. In weakness of the warrant the subject acknowledges the force of evidence, yet possesses a belief contrary to it, i.e., on Davidson's view, possesses a belief which is not caused by what counts as good grounds for it. Weakness of the will is acting where one's reasons for acting demand that one act otherwise, i.e., the cause of one's actions is not identical with one's reasons for acting.

According to Davidson this is what irrationality consists in – possessing a belief or performing an act where the mental causes of belief or action do not count as good grounds for belief or action.

Davidson recognizes that this account of irrationality raises a problem in connexion with his holistic account of the mind. For we can only identify a mental event and the belief that *p*, or the desire that *q* or the intention that *r*, in so far as it plays a role in the network of beliefs, desires, intentions and actions in which beliefs, desires and intentions count as reasons. We may be able to identify a mental event at the neurophysiological level. But at this level of description the only appropriate descriptions of connections between events are in causal terms, and not in terms of reasons. In which case we lose our hold so to speak on rationality and irrationality, and thus, at this level alone are unable to give an explanation of either. But for this reason, when we rise to the level of mentalistic description we can give a neurophysiological event its correct mentalistic description only by identifying it as a reason within a pattern of reasons and actions. This is what the holistic nature of the description of the mental life of a person amounts to; there is no piecemeal description of a mind.

For example, as I remarked above, the criteria for the correct ascription of a belief will in general be behavioural, i.e., consist of actions. If *P* performs act *a*, bringing about state-of-affairs *s*, then the criteria are satisfied (though defeasible) for ascribing to *P* the desire, *d*, that *s* and the belief, *b*, that *a* will

bring about *s*. In this case the ascription to *P* of the desire *d* and the belief *b* depend on their being *reasons* (for doing *a*).

An analogy may illuminate the point. At a physical level a certain event may be describable as one man causing the death of another. The correct description of this event at the 'social' level as the execution of Louis XVI depends on our being able to place this event and its actors in a particular framework of social and legal connections extending beyond this particular event alone. An event with an identical physical description might in one framework be an execution, in another a murder (or in another a tragic accident). Inspection of the event alone cannot tell us which; we can make sense of the suggestion that it is one or other of these only by integrating the event at the social (and mental) level into a network of social (and psychological) relations.

But if we can identify a mental event as the mental event that it is only by integrating it as a reason into the network of beliefs, desires, intentions and actions, then we are necessarily precluded from giving an account of irrationality, for irrationality was held to consist in beliefs or desires causing beliefs or actions for which they are not also reasons.

In order to circumvent this new problem Davidson points out that there are standard cases of mental events causing other mental events for which they are not reasons—viz. when each of the two events occurs in different people, i.e., 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.)[7]

Davidson's strategy is thus to account for irrationality by dividing the mind into sections which can be treated as separate like the minds of different individuals. This is how he comes to adopt the divisionist position. Note that one particular form of one person's desire causing another's belief is genuine other-deception. Clearly this account of irrationality will accommodate nicely a picture of self-deception as internalized other-deception, a picture of the Freudian sort discussed earlier. Davidson continues:

> 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. The obvious and clear cases are those of social inter-

7. D. Davidson, "Paradoxes of Irrationality," in *Philosophical Essays on Freud*, ed. Wollheim and Hopkins, Cambridge: Cambridge University Press, 1982, p. 300.

> action. But I suggest that the idea can be applied to a single mind and person. Indeed, if we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned into quasi-independent structures...[8]

This argument to divisionism is flawed in a number of places. Firstly, even if no mental event can be identified without identifiying it as a reason, this does not mean that it must be a reason for every other mental event with which it is causally linked. My desire that p may be so identified through its being a reason for a variety of intentions $I_1, I_2, \ldots, I_n$ and my belief that q may be so identified in that it is a reason for a (perhaps, though not necessarily) distinct set of intentions $I_{n+1}, I_{n+2}, \ldots I_m$. It may then be established that my desire that p and my belief that q are causally linked without the one being a reason for the other.

Secondly, crucial to the identification of a belief, desire or intention is the correct identification of which mind that belief, desire, or intention belongs to, i.e., who its possessor is. This should be simple; the possessor of a desire is that person who fulfills the behavioral criteria of that desire. And thus to distinguish different possessors and so different minds we need only distinguish between bodies. But if Davidson is to allow partitioning such that one person may possess two or more entities functioning as independent minds then this bodily/behavioral criterion of distinguishing minds is not available to him. In which case we must demand from him some alternative criterion. It might be thought that Davidson could reply that the criterion is this: that within an unique mind all the causal connections between mental events must satisfy the requirements of being reasons in the approved way. Although this will be true of Davidson's account, it cannot operate as a criterion. For the problem was one of identifiying beliefs, etc. without in every case making them conform to the standard pattern (of beliefs, etc. acting as reasons). This could be done if we can identify different minds to which to ascribe the beliefs and desires in question. In which case the correct identification of beliefs is posterior to the correct differentiation of minds. And so any criterion for the correct differentiation of minds cannot depend, as the suggested criterion does, on the prior identification of any mental events or states, etc.

Thirdly, the analogy with weakness of the will is misplaced. Weakness of the will is not a case of irrationality (at least not in the sense with which we are concerned here), for irrationality concerns relations among one's beliefs

8. Ibid p. 300.

(how they cohere and may justify one another, etc.), while weakness of the will is a matter of desires and actions. There is here an equivocation in the use of the term 'reason'. In one sense of 'reason' we can have desires from which we act which are not reasons for acting. The heroin addict may have a craving for the drug, yet he knows it will give little satisfaction and will leave him worse off than he now is. He has no reason for taking the drug, though he has a desire. On the other hand, if he does take the drug, the answer to the question 'why did he take the drug?' is that he desired it. The desire gives an explanation, i.e., a reason (in the second sense) for his so acting. Commonly we make the distinction by saying that the reason for the addict's taking the drug was that he badly wanted it, but that this was no good reason for taking it.

When Davidson says that to display weakness of will is to perform an act whose cause is not a reason for acting, he must mean that the cause is not a good reason. But this is not to say that it is not a reason which explains the act, for it may be a desire which is albeit a bad reason for acting.

For example, I may want to see a particular film tonight. But I also have a seminar paper to prepare. I know that if I go to the cinema, I will have to rise very early the next morning, which I don't like doing; I will be sleepy while working, which means I will be less effective; I will be tired for the rest of the day. At the same time, I can always see the same film tomorrow evening. I have no good reason for going to the cinema; I have every good reason for not going. Nonetheless I am weak of the will and I do go. There is however no difficulty in explaining this weak-willed action. After all I wanted to see the film. What about all my good reasons for not going? It must be that these reasons, good reasons though they were, did not constitute sufficiently strong desires to prevent me going. At the time I went to the cinema my desire to see the film was stronger than my desire to avoid rising early the next day etc. Typically, then, weakness of the will consists in acting in accord with the preponderance of my desires, when the preponderance of good reasons for acting suggest acting differently. There is no paradox, for there is no reason why my desires should match my good reasons for acting. Frequently these good reasons will concern my longer term interests, and so, when I suffer from weakness of the will, the desire to achieve these long term interests is outweighed by the stronger desire to achieve some smaller but more immediate gain.

I suggest that Davidson fails to see this because in his view all action is intentional, and because he sees having an intention as being in possession

of an aggregate of (good) reasons to act.[9] It is part of Davidson's analysis of action that it is intentional, for according to this analysis, if I Φ then Φing is an action of mine iff under some description of my Φing I Φ intentionally.[10]

The account of action is unconvincing; for it seems to make my Φing intentionally a property of the *description* of my Φing. But, of course whether or not I Φ intentionally is quite independent of any description. What different descriptions of my actions do is that some express my intentions and some do not. Even so, it is not the case that Φing is an action of mine only if I Φ intentionally. For it is quite possible to act without intention, Danto makes it clear that there are basic actions which generally do not amount to more than bodily movements, but it would be a travesty of the notion of intention to say that these are all intentional.[11] Similarly a lot of habitual or instinctive behaviour, scratching an ear lobe when thinking, wrinkling one's nose, raising eyebrows etc. constitute things that I do though without intention. (Bad habits may count as particular examples. For even on Davidson's account of intention they are not intended, for I will be able to supply no good reason for doing what I do. That acting according to the bad habit is acting is clear, not least because my habits are things I can control if I wish to). What then distinguishes what I do from what happens to me? When for instance, is a sneeze something that happens, when is it an action of mine? The most important reason for making the distinction is in order to apportion blame, praise and responsibility. We blame people for the (bad) things they do, but not for things which happen to them. If we (wrongly) blame someone for something in which they are involved, their defence may be "I couldn't help it" and this, if true, is a valid defence. If I sneeze and I genuinely couldn't help it, then the sneeze, like a heart beat, is not an action, it is something that happens to me. The defence disclaims agency by disclaiming the power to do otherwise. The mark of agency is this then: Φing is an action of mine if I could have desisted from Φing, i.e., if it was in my power not to Φ. This does not exclude acts of omission, because Φing might itself be a matter of not Ψing for some Ψ; if I could have Ψed, but did not, then Φing is an act (of omission) of mine. If I failed to prevent an accident, then this is an act for which I must accept responsibility, if I could have prevented it. If this is right then agency has no necessary element of intention; (intention may be sufficient, for it will make sense to say that I Φ intentionally only if I had a choice in Φing or not).

9. See Davidson, "Intending" in *Essays on Actions and Events*, New York: Oxford University Press, 1980, p. 85.
10. See Davidson, "Agency" in *Essays on Actions and Events*, New York: Oxford University Press, 1980, p. 50.
11. See Danto, A.C., "Basic Actions," *American Philosophical Quarterly*, 2 (1965), pp. 141–8.

Clearly there is much more to be said in the discussion of agency. Nonetheless I think there are strong grounds for my claim that not all acts involve an intention. And from this it would follow that Davidson's account of irrationality is defective. For it is part of his account that I am acting irrationally if the cause of my acting is not a reason for my acting. But in the case of actions which are unintended Davidson will be unable to identify reasons for these actions (though undoubtedly he will be able to supply causes). Consequently these actions will mark me as irrational. Yet it would seem that the bodily movements which are actions basic to all my actions must in the majority of cases fall into the category of unintentional actions, as do many actions performed out of habit or instinct. And so, on Davidson's account, most of any person's behaviour will mark them as irrational.

This suggests that Davidson's account of irrationality and weakness of the will are in error. It may nonetheless be the case that his account of irrationality is correct when restricted to beliefs, which is, of course, that part of the account which is relevant to the explanation of self-deception. This would be simply this, that it is irrational of *P* to have a belief whose cause is not also a reason or ground for holding that belief.

However, even this restricted account of the irrationality of self-deception cannot be adequate. For cases of self-deception often occur in which when I first acquire the false belief in question it does appear to be warranted, and I acquire it for all the right reasons. Self-deception comes in when I refuse to relinquish the belief once strong evidence accrues for its falsity. In this case, the false belief having been acquired for (caused by) all the right sorts of reasons, my state does not by Davidson's criteria count as self-deception.

Thus Davidson's account does not give necessary conditions for irrationality. Nor does it supply sufficient conditions. For I may acquire a belief for all the wrong reasons, e.g., a case of self-deception satisfying Davidson's account. It may turn out, by chance, that the belief acquired is in fact true. I then come to be in possession of all the relevant evidence which shows the belief to be true. Subsequently I lose the motivation for the self-deception, but being in possession of the evidence mentioned I continue to hold to my belief. Clearly at this point I am now behaving rationally, for the only reason why I now cleave to this belief is that it is warranted by the evidence. But

because the origin of this belief was of the wrong sort, by Davidson's criteria I am still in a state of irrationality.

These examples suggest where Davidson's account is in error. His account concentrates on the cause or origin of the belief in question, while what is significant to my present rationality or irrationality is not my belief's origin in the past but its present relation to acknowledged evidence.

The clear conclusion, I believe, is that Davidson's defence of the divisionist's position fails. The coexistence of incompatible active beliefs is incoherent in itself. It is also self-defeating, in that it undermines the divisionist explanation of self-deception. Analogy with weakness of the will provides no support since the analogy is false, and the account of weakness of the will mistaken in any case. Furthermore, the concomitant account of irrationality stands in need of significant amendment.

Thus I would suggest we need to avoid the divisionist account and the other-deception model in general if we are to seek a satisfactory account of self-deception, one which both avoids such problems and is true to genuine cases of self-deception.

4.

Above I raised the question: what grounds has the divisionist for ascribing incompatible beliefs? In a case of self-deception what is overt at least is this: the self-deceiver acknowledges evidence q, but asserts (and in other ways displays the belief) that not-p. However, q is excellent evidence for the proposition that p, and that not-p is false. The divisionist ascribes to the self-deceiver the (tacit) belief that p, in addition to the overt belief that not-p. The reason that the divisionist wishes to ascribe this tacit belief, if it is not merely being verbally transfixed by the other-deception model for self-deception, is the desire to maintain the rationality of the self-deceiver. And this can be supported by pointing to that fact that the self-deceiver has shown himself or herself to be perfectly rational in analogous situations. It may be that, in circumstances which involve not himself but someone else, our subject will always infer (the like of) p (the correct conclusion) from the evidence (the like of q). This shows him to be rational. And so when it comes to his own case, when he acknowledges q but denies p, the explanation that he is irrational appears unsatisfactory. It seems explanatorily more appropriate to hold that, being rational, he does reach the correct conclusion from q, that p, but that this belief is supressed.

But this argument is unconvincing; for the point with self-deception, I

claim, is that the subject is, in one particular sphere of life, being irrational. There is no inconsistency in maintaining that he or she is an intelligent and rational person who is being irrational. For this localized irrationality has an explanation, viz. a *motive.*

What counts here as irrationality is quite simple. Irrationality, in this case, amounts to having a belief in the face of evidence which shows the belief to be false. The origins of the belief in question are quite irrelelvant. What counts is that the belief is presently unsupported by anything the subject believes to be true, while its negation is well supported by acknowledged evidence. This analysis makes use of the notions of good evidence and of support between beliefs. Of course it is essential to this view that a satisfactory account of the latter would itself eschew use of the concepts of rationality or irrationality. I shall not argue the issue here, but it is my view that such an account is available. For instance, one might start with this approach: q is good evidence for p, if whenever q, then p (or at least very often p); i.e., q is a reliable indicator of p. Such an externalist account would clearly place evidence (or justification) conceptually prior to rationality.

It may be asked, what sense are we to make of the idea of a desire motivating a belief? For typically the origin of a belief is intimately connected with grounds for holding that belief. I may acquire the belief that there is a tree in the quad by seeing a tree in the quad, and this perception of a tree will count as a ground for the belief. Yet when such a belief is acquired as a consequence of a desire, the desire cannot count as a ground for holding the belief. This is important, for beliefs acquired in the former manner stand a high chance of being true and in the latter case a low chance, and it is the truth of our beliefs which guarantees the success of actions based upon them in achieving our desires. Paradigmatically, belief-desire pairs lead to action. I see that it is raining. I wish to stay dry. I believe that taking an umbrella is a good way of securing this. Therefore I take an umbrella. If my beliefs are correct I will achieve my end of not getting wet. I will maximize the chance of the truth of my beliefs by acquiring them in reliable ways: perception, good testimony, etc. I thus maximize the chances of achieving my desires. This is the way we explain intentional action.

I have suggested however that the beliefs themselves may be caused by desires. But in this case I will not be maximizing the chances of the achievement of my ends, since I will not be acquiring my beliefs by a reliable method. Imagine that all desires that p brought about the belief that p. For instance, one could, by desiring that it is not raining, thus believe that it is

not raining, and thus save oneself the effort of looking for an umbrella. Furthermore, if I do in consequence get wet, against my desire to stay dry, this is itself no problem for me, for my desire to be dry will bring about the belief that I am dry, and hence the appearance to myself at least that my desire is satisfied, and so give me the pleasure sought. But then no behaviour could be explained according to the model of belief-desire pairs leading to action, and in consequence we would deprive the concept of intentional action of any content. Not only is this a bizarre conclusion; it has simply shifted the paradox, for if we have no concept of intention, we can have no concept of motivation. Yet this conclusion derives from the premise that a belief can be motivated. And so this premise would seem to be self-defeating.

It should be apparent, however, that this reasoning must be defective. For it is clear that our desires often influence our beliefs. A trade unionist may believe that state investment in industry is the best cure for unemployment, while the city financier may believe that tax incentives are better. Although both may be irrational to let their desires or interests influence their beliefs, this phenomenon is nonetheless not uncommon. A different case of belief originating from a source unassociated with good grounds for that belief is paranoia. Here it is not the desire that *p* which causes the belief that *p*, but the fear that *p*.

The defect of the reasoning lies in this. Certainly it is nonsense to suggest that all of a person's beliefs could be the result of his or her desires. But it does not follow from this that none of the beliefs could be so caused. For although it is impossible for everyone to be taller than average, it does not follow that no one can be taller than average. Thus the occasional desire induced belief is conceivable, and quite consistent with the impossibility of the universal causing of belief by desire.

And this conclusion, that motivated irrationality must be an exception and not the rule, concurs with what has been remarked about self-deception, that it takes place against a background of rational belief formation. This would contrast with any divisionist picture of self-deception, because on the divisionist picture there would be no reason why a person could not deceive himself with respect to every belief he or she has.

Furthermore, not every desire can act as a motive for belief. For instance, *A* may deceive *B* for the pleasure of seeing him embarrass himself by being misinformed. Yet obviously I could not have this motive in deceiving myself. As Davidson's remarks do suggest, the motives for self-deception include the desire to avoid painful beliefs and to acquire pleasant ones; typically these

will concern one's own character, but they may concern someone else; the motives may indeed not be connected with any person's nature, but may aim to provide solace against some potentially painful matter of fact (one believes that *X* died peacefully, while the evidence suggests otherwise).

It would seem that the divisionist picture of self-deception would not be limited in this way. For if we are to see self-deception on the model of two parts of the mind acting as more or less independent agents, then there is no reason why one part of the mind might not deceive the other for all the reasons that agents do deceive one another, e.g., for the sake of embarrassing another person. But then the divisionist cannot explain why this does not occur in self-deception. Against this, if we take Sartre's advice and reject this divisionism, we see that the motivations for self-deception are not the motives one part of the mind has for deceiving another, but are the motives of a unitary being.

Gardiner suggests that a failing in an account such as mine is that it fails to distinguish between self-deception and wishful thinking.t[12] Pears, for instance describes wishful thinking as a belief produced by a wish. It is clear that there are similarities between self-deception and wishful-thinking. There are however clear differences. Self-deception involves greater intellectual distortion than wishful-thinking; it is stronger in two ways. Firstly there is a difference in the strength of the belief involved. The strength of a belief is related to our actions. The self-deceiver will act on the basis of his or her belief when the wishful-thinker would not. Secondly there is the strength of adherence to the belief in question. In the face of accumulating evidence against the belief the self-deceiver will continue adhere to the belief when the wishful-thinker will have already rejected it. Connected with this is the fact that wishful thinking may not necessarily be irrational. It may well be that the belief in question is true, but simply not very well supported by the evidence. (Not that all that the difference amounts to is that the belief in self-deception must be false. I may believe p despite very strong evidence that *p* is false. It may turn out, very much against the chances, that *p* is in fact true; this may nonetheless be self-deception. Someone who firmly believes that the world will end tonight is very likely deceiving himself; and that would still be the case even if some unpredictable event were to destroy the planet this evening.)

One might characterize wishful thinking as a belief where a wish induces

12. See Gardiner P. "Error, Faith and Self-Deception," in *The Philosophy of Mind*, ed. J. Glover, Oxford: Oxford University Press, 1976.

a stronger belief than is justified by the evidence. In self-deception the strength of the evidence will not simply be weak but will strongly suggest that the belief is false. Not only is self-deception stronger than wishful-thinking, but crucially it differs in content, in that self deception seems limited to the sorts of motivation indicated above (the desire for a better self-image or image of others, or the desire for solace), while the wishes present in wishful thinking will range beyond these. (One may wishfully imagine that the sun is shining. But a belief that it is so when the evidence shows it is raining would be irrational; it might indicate madness; but would it count as self-deception unless there were a powerful motive for the belief?)

Wishful thinking covers a variety of phenomena, related at greater or lesser remove from self-deception. The important features of self-deception are that the belief is irrational and that it has as its motivation particular sorts of desire. As these desires imply, and examples bear out, self-deception involves our engagement with the world in ways which are of central importance to us. Wishful thinking involves the peripheral, which is why it is weaker, why the motivation for false or undersupported belief is less powerful there. Self-deception is concerned with how we see ourselves and others, how we relate to the world and how we cope with our emotions, desires, motives. It is no wonder Sartre rejected a picture of the self-deceiver as pathologically divided; rather, unity is the essence of self-deception, its aim to preserve what is most important to us in our engagement with the world.[13]

King's College, Cambridge
Cambridge CB2 1ST, England
and
Department of Philosophy
The University of Edinburgh
Edinburgh EH8 9JX, Scotland

13. What there may be of value in this paper owes much to the insight and inspiration of Dr. John Hyman of Queen's College, Oxford.

ALEX BURRI AND STEPHAN FURRER

Truth and Knowledge of Other Minds

In a not yet published paper (??), Donald Davidson[1] makes a three-fold distinction between self-knowledge (he calls it "first person knowledge"), knowledge of the external world and knowledge of the minds of others. Although self-knowledge depends causally on the rest of knowledge, it is "logically independent of our beliefs about a world outside," as Davidson puts it, and cannot, therefore, serve as a foundation for these beliefs. The fact that we cannot be generally wrong about the contents of our own minds has led many Cartesian minded philosophers to the erroneous conclusion that the certainty of first person knowledge guarantees its usefulness in justifying external beliefs. The argument Davidson is coming to rests upon some sort of externalism. Externalism, the remaining answer to the skeptic after the failure of cartesian epistemologies, "holds that the contents of a person's propositional attitudes are partly determined by factors outside the mind," Davidson says. As "words and thoughts refer to what causes them" a linguist, who is at work on an unknown language, "cannot independently discover what an agent's beliefs are about, and then ask whether they are true." The interpretation cannot get off the ground if the linguist refuses to treat the agent as a speaker of truth, i.e., if he eliminates the so-called principle of charity.

In our opinion, Davidson's three-fold distinction helps to elucidate the role truth plays in acquisition of knowledge, but his own account of "truth mechanics" is bound to fail. We hope to clarify our point as we go along. Our concern is with the relationship between knowledge and truth. If knowledge is identified with justified true belief, then this relationship, one may assume, turns out to be merely terminological, hence uninteresting: knowledge always is true, whereas beliefs can also be false; therefore, the relevant epistemological question is whether a single belief (or a whole body of beliefs)

1. "Epistemology Externalized," forthcoming in *Dialectica* 45 (1991).