An Analysis of Self-Deception

Kent Bach

Philosophy and Phenomenological Research, Volume 41, Issue 3 (Mar., 1981), 351-370.

Stable URL:
http://links.jstor.org/sici?sici=0031-8205%28198103%2941%3A3%3C351%3AAAOS%3E2.0.CO%3B2-K

AN ANALYSIS OF SELF-DECEPTION

Many puzzles surround the topic of self-deception. What distinguishes cases of self-deception from such allied psychological phenomena as wishful thinking, intellectual blindness, biased thinking, and other forms of irrationality? What are the distinguishing features of self-deception, or is there such a diversity of cases that no unitary analysis can be given? Is 'self-deception' an accurate or at least not a misleading description of the phenomena so labeled? Behind these questions, themselves amply difficult, lurk the seeming paradoxes of self-deception. Self-deception seems to be, if it is what the name implies, the schizoid act of directly and knowingly getting oneself to believe what one disbelieves (or to disbelieve what one believes). Thus it is often described, by philosophers and others as well, as 'lying to oneself' or as 'fooling oneself.' The supposed paradox consists not merely in the victim's (or is he the benefactor?) simultaneously holding contradictory beliefs or even in his awareness of this fact, but also in the apparent intentionality with which he accomplishes his feat. The self-deceiver does not hire a hypnotist or summon a brainwasher to do the job—he does it himself. What makes his accomplishment so perplexing is that he must know what he is doing if he is to do it and yet not know if it is to have effect. Or so it seems. Perhaps a different conception of self-deception is in order.

I. WHAT SELF-DECEPTION IS NOT

To demarcate our topic let us mention several phenomena that are distinct from self-deception (although it may include these at times). (1) It is not wishful thinking or if it is, it is a very special case. Wishful thinking need not involve any reasoning or semblance of reasoning. The wishful thinker imagines some state of affairs, likes what he imagines, and supposes that it will transpire. He does not try to justify this supposition, perhaps being content with the absence of evidence one way or the other. Were he aware of counterevidence and of the need to deal with it, we would have the special case of self-deception. (2) Nor is self-deception simply a case of intellectual blindness. That consists in failing to see where the evidence or the reasons point, whereas the self-deceiver sees this all too well, at least at the

outset. (3) Although self-deception may serve a person's interests, it is not simply a case of biased thinking. When we charge someone with bias or prejudice, we imply that his thinking is adversely affected by his sentiments, which render it peculiarly inflexible, but we do not imply either that any special effort is being made (bigotry can be effortless) or that there is something *uncharacteristically* irrational in the person's thinking. Self-deception does seem to have these features. (Finding his intellectual antics unfathomable, we may at times accuse the bigot of self-deception, but this is to credit him with more than he deserves.) (4) Self-deception is no ordinary case of uncharacteristically irrational thinking, which could be due to fatigue, shock, confusion, alcohol, or whatever. A fit of irrationality need not be self-servingly motivated but self-deception is, in a specific way to be explained in due course.

Various accounts of self-deception have been proposed, attempting to show either that it is not paradoxical after all or that there is no such thing. To show that it is not paradoxical people have appealed to such notions as unconscious intentions, half beliefs, and multiple selves, notions which if not themselves paradoxical are surely problematic. Others have suggested that the air of paradox is produced by the label, 'self-deception,' and that what we misleadingly call by that name can be assimilated to other, less puzzling psychological phenomena. Herbert Fingarette devotes a chapter of his book *Self-Deception*[1] to these various proposals. He persuasively argues that in each case either there is a residue of paradox or paradox is eliminated only by leaving out something essential to self-deception. He holds that what these various analyses neglect is not what the self-deceiver believes but what he does. In later chapters Fingarette offers fascinating speculations about what the self-deceiver does, but no analysis is presented and it is not clear how an analysis could be extracted from these speculations.[2]

David Pears[3] has since attempted to resolve the paradoxes of self-

[1] New York, Humanities Press, 1969, Chapter II.

[2] Fingarette does say that the self-deceiver 'persistently avoids spelling out some feature of his engagement in the world' (*op. cit.*, p. 47), but it is not obvious how to construe this as an analysis, especially because there is no explicit indication of what sort of proposition the self-deception is supposed to be about. Besides, as D. W. Hamlyn has pointed out, Fingarette overlooks the case in which the self-deceiver is *too* intent on spelling out his engagement (*Proceedings of the Aristotelian Society*, Supp. Vol, 1971, pp. 50-51).

[3] 'Freud, Sartre, and Self-Deception,' in *Freud*, edited by Richard Wollheim (New York, Doubleday, 1974), pp. 97-112.

deception in a way immune to Fingarette's criticisms of previous efforts. Pears observes that what makes self-deception seem paradoxical is not merely what the self-deceiver believes or the irrationality of it but also what the self-deceiver does which, according to Pears, is intentionally cultivating an irrational belief.

In order to remove the appearance of incoherence from the self-deceiver's plan to instill in himself a belief contrary to what he already believes or has reason to believe, Pears suggests that the self-deceiver, though quite aware of his 'rational tendency' to believe not-p, can rely on his wish to believe that p. In embarking on a series of measures designed to lead him to believe this, the conscious self-deceiver must take care that along the way he loses awareness of why he is engaging in this project. He must be aware at the outset that his wish that p is needed to produce uncharacteristic distortion in his mental processes, but part of his plan is to rid himself of this awareness later. Pears points out that this plan appears incoherent because, as the person becomes less and less aware of his motive and of how he is to fulfill it, seemingly it will become less and less effective and the plan will fail. However, Pears sees no reason to deny that the self-deceiver, whose plan has the peculiar feature that it 'cannot be fully reviewed when it approaches completion' (107), can rely 'on the discreet operation of his wish to believe p' (109). Pears realizes that how all this works needs explanation, but he sees nothing incoherent in the idea of the self-deceiver's relying on his wish to believe that p and thereby coming to believe it.

The trouble with Pears' account is that it is based on a misdescription of the self-deceiver's wish as 'the wish to produce the belief that p' (p. 104), whereas his real concern is with what is the case, not his belief about it. This error is the same as the common misdescription of wishful thinking as 'believing what you want to believe.' Surely what the wishful thinker believes is what he wants to be so. Quite different is, e.g., the case of Pascal who, though doubtful of God's existence, got himself to believe in God not because he wanted God to exist but because, believing that if God exists He sends to heaven only those who believe in Him, he wanted to be on the safe side. If Pascal did not care whether God existed, he would not have believed what he wanted to be the case but only what we wanted to believe. If the wishful thinker merely wanted to cause himself to believe something, he would not care whether his belief were true, but surely he does, as does the self-deceiver.

As I will propose, self-deception is not essentially a matter of

belief at all. A person who believes that *p* (or that the evidence heavily favors *p*) can deceive himself that not-*p* without having to get himself to believe that not-*p*. Consider that the occasion for deceiving oneself arises only insofar as the touchy subject is thought of, and so if the person believed that *p* (while desiring that not-*p*) but it never occurred to him that *p*, he would have no occasion to deceive himself. Accordingly, what matters in self-deception is not the belief that *p per se* but the occurrence of the thought that *p*, especially on a sustained or repeated basis.

On the view to be proposed here, the self-deceiver desires that not-*p* while believing that *p*, and what he does is to avoid the sustained or recurrent thought that *p*. As we will see, there are three distinct ways of doing this, none of which involves paradox. Before describing them and extracting an analysis of self-deception from what they have in common, I need to explain the important distinction on which I rely, the distinction between believing and thinking that *p*.

II. THINKING AND BELIEVING

Philosophers sometimes distinguish between occurrent and dispositional senses of 'believe,' but I will use the term 'believe' only for the dispositional sense and reserve the word 'think' for the would-be occurrent sense. I say 'would-be' because I deny that occurrent believing is believing at all, or in my terminology, that thinking that *p* is either necessary or sufficient for believing that *p*. Surely it is not necessary, for not only do we all have countless beliefs that we are not currently entertaining, we have many whose content we have never had in mind, e.g., (until now), that kangaroos are bigger than cockatoos and that the moon is not made of bleu cheese. I claim also, perhaps controversially, that thinking that *p* is not sufficient for believing that *p*, although one generally thinks that *p* only if one believes that *p*. I can best defend this claim after making some further points about thinking and believing.

I distinguish between thinking *that p* and thinking *of p*. One can think of *p* without thinking that *p*, for one might think that not-*p* or not know what to think about it. Thus, thinking that *p*, thinking that not-*p*, and not thinking either are compatible with thinking of *p*. Whereas both thinking of and thinking that are mental occurrences, thinking involves at least momentarily assenting to or judging that *p*. This, I claim, does not entail believing that *p*, even for a moment.

Unlike thoughts beliefs are states, not occurrences. Consider the currently popular view of beliefs as functional states. The term 'functional,' as applied not only to beliefs but also to desires, intentions, and emotions, implies that a state must be characterized by its role in a system of such states. Unlike the older, behaviorist conception, the functional approach requires that psychological states be characterized not merely by their relations to sensory stimulation ('input') and to behavior ('output') but also by their relations to each other. Moreover, functional states such as beliefs are representational, having propositional content. Accordingly, a well-developed individual psychology couched in this framework would contain systematic generalizations about the relations among a person's states and about how they mediate his experience and what he does.[4]

This capsule statement of functionalism obviously needs elaboration, and it is hoped that some hearty functionalist will someday spell out the details. I do not rely on the correctness of this conception, for my aim is only to contrast beliefs with thinkings-that, and for this even a simplistic behaviorist analysis will do. It is enough that beliefs not be confused with occurrences of the corresponding thoughts, i.e., the belief that p with the thought that p. By recognizing this distinction we can allow not only for the obvious fact that a person need not have thought everything he believes but also for the fact that his thoughts do not invariably correspond to his beliefs and that they are not invariably what he thinks they are. This is not to deny, of course, that usually what a person thinks, and thinks he believes, is in fact what he believes. Indeed, a functionalist account would be incomplete if, in addition to characterizing a given belief in terms of its relation to a person's other beliefs, it did not recognize that, at least in general, if he believes that p, then whenever he thinks of p he thinks that p. In other words, to believe that p is to know, should the thought of p occur, what to think about it without having to deliberate. This is 'knowing what' in the practical sense of knowing what to do (or think) in a given situation. It does not entail that what one knows to do is the right thing, nor does it imply that one will do it. One might get flustered or, for that matter, not act automatically but reconsider what to do.

As Davidson has insisted,[5] because there are many ways a given

[4] Gilbert Harman formulates these features of the functionalist position in *Thought* (Princeton, N. J., Princeton University Press, 1973), pp. 43-46 and 62-65.

[5] See, e.g., Donald Davidson, 'Thought and Talk,' in *Mind & Language*, edited by Samuel Guttenplan (Oxford, Oxford University Press, 1975), pp. 7-23.

belief can link up with others and with behavior, no single condition is necessary or sufficient for a person's having the belief (of course, if it did not link up somehow, there would be no basis for ascribing it). Without making assumptions about the person's other beliefs and without imputing some degree of rationality to him (not that these assumptions should be fixed), we could not justifiably ascribe the belief in question. I should add that we cannot assume a person always to believe what he thinks he believes, for then we would be arbitrarily ruling out the possibility of error about one's beliefs. Now consider thinkings-that. They are not states but occurrences, and need not be supposed integrated with beliefs or with each other to be justifiably imputed to someone. We can best rely on his (sincere) testimony, on what he thinks he thinks (even if we do not assume incorrigibility, the possibility of error about what one currently thinks is much narrower than about what one believes). Thus, however strong the evidence that someone does not now or later believe that p, relative to his (sincere) testimony this is weak evidence that he does not think that p. Without having believed that p he could now think that p without coming to believe it. The thought that p could admittedly be a passing fancy.

Here it might be objected that thinking that p is at least believing that p at the time, even though the belief does not 'take.' I do not wish to exclude the possibility of short-term beliefs (perceptual beliefs are an obvious example, though they stem not from thinkings-that but from perceivings-that), but consider the following case. Someone thinks that p for a moment and then notices an unacceptable implication of it. He therefore ceases to think that p. Did he believe p for that moment? I would describe him not as giving up a momentary belief but as keeping himself from believing that p.

Perhaps no such example or any argument will keep someone who insists from calling thinkings-that 'occurrent beliefs,' but if more is at stake than mere terminology, he must offer an alternate conception of belief that justifies describing thought occurrences as beliefs. In any case, I hope the following examples help clarify the distinction between thinking (-that) and believing. Already mentioned was the case of someone on a whim momentarily thinking that p without forming a belief to that effect. Similarly, one might be momentarily deceived by a loud bang into thinking a gun had been fired or tempted by a piece of flattery into thinking one's lost youth had returned, without forming the corresponding belief. Vacillation provides another illustration. Suppose someone is presented with persuasive

arguments for and against *p*. Taken in for the moment by the first he thinks that *p* and then, even more impressed with the second, he thinks that not-*p*. He may then reconsider the first argument, still finding it compelling, and proceed to alternate between thinking that *p* and thinking that not-*p* (what he should realize is that he does not know what to think). Since no position has been settled on, we should not describe this as a case of alternatingly believing and disbelieving that *p*. A final illustration of how believing and thinking can pull apart is provided by phobias. Consider someone who, whenever the prospect of traveling by air arises, develops acute anxiety about flying even though he believes that commercial flying is as safe as many other things he does without fear. Even while realizing the irrationality of it, he cannot help thinking that flying is dangerous.

Much more could be said to clarify the contrast between thinking and believing and to support the claim that thinking does not imply believing. For present purposes anyone who insists on retaining the occurrent as well as the dispositional sense of 'believe' can substitute in what follows 'occurrently believe that *p*' for 'think that *p*' and 'dispositionally believe that *p*' for 'believe that *p*.' Though not put as neatly, the underlying point of the proposed analysis of self-deception will be preserved, that what matters is what *occurs* to the self-deceiver, not what he believes in the full dispositional sense.

III. THREE WAYS OF DECEIVING ONESELF

The self-deceiver, believing that *p* while desiring that not-*p*, need not, on my view, try to get himself to believe that not-*p*. That is neither his objective nor essential to it. It is enough that he not (sustainedly or repeatedly) think what he believes, for what matters is what occurs to him. Self-deception need not (though it can) lead to change in belief—it is the thought that counts.

There are three basic ways of avoiding the thought that *p*, which I dub *rationalization, evasion,* and *jamming*. These are activities or processes taking place at definite times. Distinct from the process of deceiving oneself, of which these are the basic forms, is the state of being self-deceived, which results from this process. The process of deceiving oneself about *p* occurs only when *p* comes to mind, whereas one can be in the state of being self-deceived about *p* even when not thinking of it. However, to be in that state means that were one to think of *p*, one would avoid the thought that *p*. Indeed, evasion and

jamming, which (as we will see) are not as elaborate as rationalization, are well-suited not only for deceiving oneself in the first place but also for maintaining that state on occasions when the thought of *p* occurs. The analysis proposed in the next section will be extrapolated from our description of the three ways of deceiving oneself (or remaining self-deceived), but it will be formulated to characterize the state of being self-deceived itself.

Rationalization

In psychological contexts rationalization is understood as a person's makeshift justification of an action in terms of motives that seem to others not to be his genuine ones. Insofar as the person is sincere about this justification, we are inclined to regard him as having deceived himself. For our purposes we will not restrict use of the term 'rationalization' to self-ascription of motives, and will extend it to cover any case of a person's explaining away what he would normally regard as adequate evidence for a certain proposition. This might, but need not, be a proposition about an issue on which he already has a contrary belief, a belief now in jeopardy from new evidence to the contrary.

There is nothing intrinsically irrational about explaining away evidence against what we already believe. This is part and parcel of good scientific method—theories should not be discarded without a fight—and of everyday thinking as well. Initially we try to deal with 'recalcitrant experiences' not by adjusting our beliefs but by looking for something wrong with the experiences. Before making serious changes in our beliefs, we try to render contrary data not worthy of accommodation. Only if we cannot do this without constructing theoretical epicycles do we adjust our beliefs, making changes as local and minimal as possible. There seems to be no hard and fast boundary between doing this rationally and not. Dealing with recalcitrant data is always a matter of give and take between what we believe already and our current experience, and while it is reasonable to adjust what we believe to accommodate new data, it is also reasonable to reject data because we cannot accommodate them, especially if they can be explained away consistently with what we believe already. But there are limits, even if it is not clear where they are to be drawn.

There *is* something intrinsically irrational about explaining away evidence because it weighs against what one desires, especially if, but for that desire, one would adjust one's belief to the evidence. Rationalization in self-deception sometimes knows no limits. A classic

example is the case of the man who believes a certain woman loves him even though he possesses (and realizes he possesses) strong evidence to the contrary. He is cognitively unmoved by the fact that she has never wanted to go out with him, always hangs up on him whenever he calls, returns his unsolicited gifts, plans to marry someone else, etc. He 'knows' there must be some explanation for all this: her mother poisoned her mind against him, her father wants her to marry someone respectable, she thinks he is too good for her, she does not realize how much he loves her, or whatever. The rationalizer does not disregard the evidence against what he desires but explains it away by constructing hypotheses that render it compatible with what he desires. These hypotheses may seem wild to us but not to him.

It is not the irrationality of the rationalizing self-deceiver that we find so puzzling. We might be somewhat puzzled by a person who was consistently irrational (or whose standards of rationality, though consistently followed, were fundamentally different from ours). The self-deceiver's irrationality is especially puzzling—and distinctive—because (1) it violates *his* standards of rationality, (2) this violation is uncharacteristic of him, and (3) he denies its existence. Of course, for him to acknowledge his irrationality would tend to undermine it and its (to us) obvious purpose, which is to make things seem to be as he wants them to be, a purpose he cannot coherently avow, at least once it is achieved.

The rationalizer would normally believe the evidence adequate and would not invent hypotheses to explain it away. However, in cases like the above, he believes the evidence to be inadequate, even if he himself is aware that he would normally find it adequate. Somehow in this case he is convinced that something is wrong with it or missing from it, even if he cannot say what. The wholesale explaining away characteristic of rationalization is not a matter of explicit policy. In particular, the rationalizer need not believe that no body of evidence, no matter how overwhelming, could be adequate for p (unless be believes that p could not possibly be true). Rather, he believes that the evidence he is presented with is inadequate. It may be true that he would believe this of any body of evidence with which he might be presented, but that does not mean he has a general policy to that effect.

So far we have looked at the 'reasoning' of the rationalizer, on the basis of which he denies that p. We should not assume that he thereby believes that not-p. He may believe that not-p—indeed, he may very well have believed it all along—but for him to deceive

himself, it is enough that he think that not-p at the time. Reasoning to a conclusion need not lead to belief, for the conclusion may not 'take.' Even if one does believe the conclusion, it need not be the reasoning that led to the belief, and one may be mistaken in so thinking. This is common in rationalization, inasmuch as the rationalizer, if he does believe that not-p, would probably continue to believe it with or without the reasoning. His 'reasoning' does not really lead to the belief but serves, rather, to convince him that he is justified in believing that not-p. He may, on the other hand, have no belief about p at the time, but repeated rationalization may affect not only what he thinks about p when the issue comes up but also what he actually believes about p.

Rationalization about a particular proposition can lead to rationalization about other, more general ones, e.g., that one's senses are reliable or that others' testimony can be trusted. The reverse can also happen and often does. For example, if a parent believes his child to be a 'good little boy,' he may find it difficult to believe the child could have kicked the neighbor's dog. Kicking dogs is bad, and therefore 'Johnny could never have done a thing like that.' Similarly, if a parent believes he loves his child, then when he strikes out at the child for making too much noise, he reasons that he did it not out of rage but 'for Johnny's own good.' In cases like these, the self-deceiving rationalizer does not merely desire that p not be true, he believes a generalization that is incompatible with p and reasons accordingly.

Evasion

Turning one's attention away from some touchy subject is what I call 'evasion.' This is a common phenomenon and is generally not self-deceptive in character. For example, one might keep one's mind off some embarrassing episode or some harrowing experience without deceiving oneself about it, simply because one does not wish to be reminded of it. One may have developed the habit of turning one's attention to something else whenever the touchy subject comes to mind. This is much like the technique of distracting oneself to reduce awareness of pain or nausea.

Whereas the simple evader admits what he believes (being resigned to it) but would rather think of something else, the self-deceptive evader avoids the thought *of* p specifically to avoid the thought *that* p. The obvious way to do this is to think of some consideration against p. This may not be an especially strong consideration, even by his own standards, but it is strong enough (psycho-

logically, if not epistemologically) to get his mind off p. Doing this, evasion might seem like rationalization (in one step), inasmuch as one thinks of a reason against thinking that p, but it is really not even that. The evader does not assess the strength of this reason or evaluate it against reasons to the contrary, and does not explicitly conclude that not-p. Rather, he merely thinks of a single reason against p and turns his mind to something else. It does not matter how good the reason is, even by his standards, but only that he does not think any further on the issue (if he does think further, rationalization may take place). Such self-deceptive evasion is to thought as procrastination is to action. In procrastination one avoids action by thinking of a reason against it, however weak that reason may be, and then turns one's attention to something else.

Whereas rationalization lends itself to deceiving oneself in the first place, typically when one is confronted with strong evidence against what one wants to be so, evasion best serves to preserve the state of being self-deceived. Recall that generally if a person believes that p, then whenever he thinks of p he thinks that p. However, a person who is self-deceived that not-p has a competing disposition not to think that p despite believing it. This disposition may have been instituted by an original rationalization, even if that rationalization did not lead to his believing that not-p instead of p, but it does not require rehearsal of that rationalization to be effective. Should the self-deceived person happen to think of p, he may succeed in avoiding the thought that p just by thinking of one reason, perhaps recalled from the original rationalization, against it. Then again, this reason might be newly contrived. More efficient is simply the thought that there is some such reason—identifying it is unnecessary—and that itself serves as a reason. Most efficient is the thought that p is not worth thinking about. What better reason not to think about it?

Jamming

A person can believe that p without thinking that p. Even if he cannot avoid thinking of p, he can, whenever p occurs to him, think that not-p. Whenever the issue of p comes up, as a result of his desire that not-p, he focuses his attention on what it would be like if not-p were true and vividly imagines desirable consequences of not-p. He may run through evidence favoring not-p and perhaps even go so far as to provide himself with instant 'evidence' for not-p, e.g., by acting as if not-p were the case or by convincing others of not-p and then taking their word for it. By such means as these, whenever the

thought of p occurs to him, he can 'jam' his belief that p and think that not-p instead.

Jamming is particularly effective in self-deception about one's feelings or motives. For example, a person might resent having to care for his aged mother and wish she were dead. He might believe that he feels this way, and the evidence that he believes this, although he never admits it, is that the thought of feeling this way frequently occurs to him. However, he never thinks *that* he feels this way. Instead, whenever the subject comes up (his mother often brings it up), he thinks nice thoughts about her and does something special for her, such as buying her roses and showing her old photographs. He is especially good to her when other people are around, who later praise him for his devotion. In this way, whenever the issue arises, he jams his belief about how he really feels, thinking that he loves his mother dearly.

Like self-deceptive evasion, jamming occurs primarily in the service of an already established state of self-deception. Neither presents the semblance of full-blown reasoning that rationalization does, but both contain elements of reasoning. Whereas evasion provides a consideration against p psychologically sufficient to get one's mind off p, jamming clutters one's mind with considerations against p or favoring not-p. Of course, since the jammer believes that p (or has what he would normally regard as good reason for believing that p), jamming is a highly selective process, focussing primarily on reasons to the contrary and on possible defects with reasons in favor of p.

IV. THE ANALYSIS

The three ways of deceiving oneself or staying self-deceived are distinguished by what the self-deceiver does to avoid the sustained or recurrent thought that not-p. Despite their differences, what rationalization, evasion, and jamming have in common is the state they yield or preserve, that of being self-deceived. The following is a first approximation of what it is for a person S to be self-deceived (that not-p) over a period of time t_1-t_2:

Over t_1-t_2 is self-deceived that not-p if and only if, over t_1-t_2,

(1) S desires that not-p,[6]

[6] Some cases might be more aptly (and specifically) described in terms of an emotion. For example, a self-deceiver might dread that p and thereby desire that not-p. In our later discussion of how the self-deceiver's desire contributes to his motivation, it should be kept in mind that an emotion might underlie that desire and therefore be the fundamental motivating factor.

(2) *S* believes that *p* (or that he has strong evidence for *p*), and
(3) (1) and (2) combine to cause *S* to avoid the sustained or recurrent thought that *p*.

As it stands, this analysis is too weak, for it lets in cases of evasion that are not self-deceptive. Consider the case of a man who (1) desires his estranged wife to return to him, (2) believes that she never will, and (3) as a result of his desire and belief, destroys all reminders of his wife, takes up with another woman, and does whatever else is necessary (short of committing suicide or undergoing a lobotomy) to avoid thinking about her at all. He thereby avoids the sustained or recurrent thought that she will not return to him, but clearly he has not deceived himself that she will not—he merely wants not to think about it. Clearly the reason this example does not count as self-deception is that the person avoids the thought *of p* in its own right, only incidentally avoiding the thought *that p*.

How can our analysis be strengthened to exclude the case of the evader who wants merely not to think about the touchy subject? Suppose we added the conditional, 'whenever the thought of *p* occurs to him, he avoids the sustained thought that *p*.' The trouble is that this conditional is as true of the simple evader as of the self-deceptive one, because both avoid the sustained thought *of p*. Besides, being indicative it could be vacuously true if the evader never thinks of *p* at all. But now consider what *would* be true of evaders of each kind if, contrary to fact, the thought of *p* were to occur to them on a sustained or recurrent basis. The simple evader would resignedly face up to the fact that *p*, whereas the self-deceptive evader, being unable to avoid the thought *of p*, would do something else to avoid the thought *that p*, i.e., resort to jamming or rationalization; or, if he did nothing, he would become undeceived. Thus, the following subjunctive conditional is true only of the self-deceptive evader, that even if the sustained or recurrent thought of *p* were to occur to him, he would still avoid the sustained (or recurrent) thought that *p*. So let us add a fourth condition to our analysis:

(4) If (3) is satisfied by *S*'s avoiding the sustained or recurrent thought *of p*, then even if the sustained or recurrent thought of *p* were to occur to him during t_1-t_2, (1) and (2) would still cause *S* to avoid the sustained or recurrent thought that *p*.

There is another problem for our analysis. It is analogous to the problem of 'wayward causal chains' in connection with causal theories of action, perception, and reference. Suppose the conditions of our analysis are satisfied, but S's desire and belief do not have their effect

in the right way, the way peculiar to self-deception. Imagine that I desire not to have the flu (say I have an important interview coming up) but that, given certain symptons, I believe I have adequate evidence that I do have the flu. However, this desire and this belief combine to cause me to feel great anxiety. Feeling the anxiety I think that my symptons are produced by the anxiety, not by the flu, Clearly I am not engaged in self-deception. My desire and belief cause me not to think that I have the flu, but they have this effect in the wrong way. What is the right way?

As formulated above, our analysis of self-deception leaves open what kind of causal connection S's desire and belief have to his avoidance of the thought that p. It might be proposed that his desire and belief jointly constitute a *reason* (let us assume that reasons are causes) to avoid the thought that p. As the various examples in section III make abundantly clear, being self-deceived about p is compatible with believing that p, provided one does not think that p (at least on a sustained or recurrent basis) when the thought of p occurs. However, it is difficult to conceive of how the belief that p together with the desire that not-p could constitute a reason he could act on. Surely he does not reason, 'Although I believe that p, since I desire that not-p I will avoid the thought that p whenever p comes to mind.' If that were what self-deception involved, it would be intolerably paradoxical.

Although the self-deceiver's desire and belief do not constitute a reason on which he acts to avoid the thought that p, still they combine to *motivate* him to avoid that thought. In the flu case, which shows that a merely causal analysis will not do, the person's desire not to have the flu and his belief that he has adequate evidence for his having the flu cause him, by way of producing anxiety that seems to be the cause of his symptons, not to think he has the flu. However, they do not *motivate* him to avoid thinking that he has the flu. Let us require, then, that the self-deceiver's desire and belief combine to motivate him to avoid the thought that p. We still need to explain just what motivation is, but first we may offer the following as our final analysis of self-deception:

Over t_1-t_2 S is self-deceived that not-p if and only if, over t_1-t_2,

(1) S desires that not-p,

(2) S believes that p (or that he has strong evidence for p),

(3) (1) and (2) combine to motivate him to avoid, and he does avoid, the sustained or recurrent thought that p, and

(4) if (3) is satisfied by S's avoiding the sustained or recurrent

thought *of* p, then even if the sustained or recurrent thought of p were to occur to him during t_1-t_2, (1) and (2) would still motivate S to avoid, and he would still avoid, the sustained or recurrent thought that p.

These conditions cannot be satisfied by the counterexamples to our original formulation: (3) now requires that (1) and (2) not just cause but motivates S's avoidance of the thought that p, and (4) excludes the case of the simple evader, whose motivation would not have effect if he could not avoid the thought of p.

V. THE SELF-DECEIVER'S MOTIVATION

What is it for the self-deceiver's desire that not-p and his belief that p (or that he has strong evidence for p) jointly to motivate his avoidance of the (sustained or recurrent) thought that p? To answer this question we must first explain the notion of motivation being used. Then we must examine how the self-deceiver is motivated to do what he does and how this motivation is tied to the intentions with which he does it. After that we will be in a position to consider whether there is anything paradoxical or unintelligible about self-deception.

Without surveying the diverse ways in which the concept of motivation has been used, I will simply explain my use of it in the above analysis. Although I know of no explicit precedent for my formulation, I believe (but will not here show) it to be implicit in much psychological discourse. It seems to me that motivation is a special kind of psychological causation, in which the relevant cause is a psychological state or combination of states, such as beliefs, desires, or emotions. Now when we say that someone is motivated (e.g., by love, curiosity, or fear) to do something, we do not imply that he will do it but at least that he is inclined to do it. In particular, I say, it motivates him to do something if it *causes him to accept a reason* for that action. In some cases it may even cause him to think of the reason. Commonly the content of the motivating state constitutes the reason, but this need not be so and is not in the case of many interesting psychological phenomena. For example, a child's desire for attention might motivate him to complain to his parents. His complaint is not lack of attention but something else, and yet lack of attention is what causes this other matter to be what he complains about. That is why, in such cases, we are inclined to say that if he did not complain about that, he would have complained about some-

thing else. Another illustration is the malingerer, whose desire (say) to avoid responsibility causes him to take ill health as a reason for staying in bed. Compulsive behavior is another common example. In each case the person does something for a certain reason that he is willing to acknowledge but, as we sometimes put it (because we find the professed reason hard to swallow), his 'real reason' is something else, something that he disavows in all sincerity. Rather than call it a reason at all, I prefer to call it his motivation and to regard his avowed reason as the reason on which he acts, even though he would not act on that reason but for the underlying motivation.

As stipulated by our analysis, the self-deceiver's desire and belief combine to motivate him to avoid the (sustained or recurrent) thought that p. It is true that he cannot bear the thought that p, but this is not his reason for avoiding it. He has his reason for avoiding it, and that is the reason that implements the process of rationalization, evasion, or jamming as the case may be. All three of these processes involve, to different degrees, the thought of reasons for rejecting p or reasons in favor of accepting not-p. The self-deceiver's desire that not-p and his belief that p combine to cause him to accept reasons for avoiding the thought that p.

In construing self-deception as a special case of the common phenomenon of being caused to accept reasons to do or think something, we should not overlook its special features. In particular, deceiving oneself through rationalization can be a long and involved process, and so when we say that the self-deceiver's desire and belief cause him to accept a reason to avoid the thought that p, we must recognize that the precise way in which he arrives at this reason is not so caused. Even though he is motivated to reach the conclusion he does, just how requires further explanation. However, we may reasonably suppose, considering his motivation, that he would have reached this conclusion somehow.

Another characteristic feature of the self-deceiver is his innocent denial of what he is doing and of course, of his motivation for doing it. Even when confronted with the charge that his dealings with p flout his own rational standards, he may sincerely deny this. Instead, he may engage in further rationalization, this time trying to convince his accuser and ultimately deceiving at least himself about his reasoning on p, as well as about p itself. The fact that the self-deceiver's motivation may be strong enough to support not only the original self-deception but subsequent defenses of it (the 'buoyancy of his

wish,' as Pears calls it[7]) seems to require further explanation. Let us consider separately (1) the self-deceiver's lack of awareness (when not accused) of violating his own rational standards and (2) his sincere denial of the charge of such violation, for whereas an accusation has to be met, unawareness is merely an omission.

(1) It might seem that the self-deceiver is somehow prevented from becoming aware of his violation (or of his belief that p), inasmuch as such awareness would vitiate the process of deceiving himself in the first place or would undo it later. The possibility of repression might be invoked here. However, although this may be a genuine phenomenon, we should not use it to explain the self-deceiver's lack of awareness. Were his unawareness due to repression, his avoidance of the thought that p would be explained but would not be a case of self-deception. The self-deceiver is not forced to avoid the thought that p, only motivated to avoid it. Instead of invoking repression, let us reconsider the supposition that the self-deceiver is prevented from being aware of violating his rational standards (or of his belief that p). Perhaps it only seems that way. After all, that such awareness would vitiate or undo his self-deception proves nothing, for there is no reason to assume that he must deceive himself. Either he has this awareness or he does not. If he has it, he will not be able to deceive himself or (if already self-deceived) he will become undeceived, but people often fail to deceive themselves or become undeceived. If he lacks this awareness, then his desire that not-p will combine with his belief that p to motivate him to avoid the thought that p, and he will avoid the thought that p. When in the future he again thinks of p or thinks of considerations bearing on the truth or falsity of p, if he is still so motivated he may continue to deceive himself about p. Or he may not. Self-deception is not inevitable, though it may seem that way.

(2) When the self-deceiver is confronted with what he is doing and denies it, he seems doubly motivated: not only to deceive himself about p but also to be unaware of what he is doing (perhaps even deceiving himself about that). Again we need not posit repression to account for what happens. What happens is that the self-deceiver, even when confronted, lacks some awareness at some point. If he admits the weakness of his earlier reasoning, he may bolster it with further considerations but not recognize their weakness in turn. No matter how far his accuser (friend, psychiatrist) goes and no matter how

[7] *Op. cit.*, p. 110.

much is conceded, there is always room for further considerations or reconsideration of previous ones. His accuser may call this 'resistance,' but as long as there is something about what he is doing whose violation of his own rational standards he does not notice, he *can* continue to deceive himself. This does not mean that he will. He is motivated to avoid the thought that p, but this does not mean he is motivated to go to endless lengths.

We should be careful, then, not to try to explain too much by the self-deceiver's motivation. Only what the self-deceiver does needs to be explained, not what he might do if he were confronted with further considerations or with accusations about what he has done already. Any subsequent disavowals or further self-deceptions require further explanation. Although the self-deceiver may fortify his position against any subsequent challenge, there being no limit to the intellectual contortions he might perform, we cannot assume that he is presently prepared to do all this. For all we know now, something might very well render him undeceived later.

Even though the self-deceiver is not fully aware of what he is doing, we do hold him responsible for it. My view of self-deception might seem to undermine any ascription of responsibility, since it denies that self-deception is intentional. That is, although the self-deceiver does what he does intentionally, he does not do it under the description of 'deceiving myself' or anything of the sort. Rather, he is motivated to avoid the thought that p but is unaware of (or denies the impact of) this motivation and of his uncharacteristic violation of his own rational standards. However, one can be responsible for something without having done it intentionally. Negligence is a prime example of this, and that is precisely what the self-deceiver is guilty of. He is not as careful or attentive as usual, he does not guard against the influence of his desires, he wants to avoid the issue. The power of his motivation does not automatically excuse his negligence.

VI. CONCLUSION

The account proposed here attempts to capture the complexities of self-deception without the paradoxes. The key is to distinguish what the self-deceiver thinks, when the touchy subject comes to mind, from what he believes. Normally the state of believing that p causes one to think that p whenever the thought of p occurs, but the state of being self-deceived overcomes this tendency. The three ways of deceiving oneself or of remaining self-deceived — rationalization,

evasion, and jamming—involve intentional action, but the content of the self-deceiver's intention is not to deceive himself or to violate his own rational standards, even though this is the effect of what he does. What he does—avoiding the sustained or recurrent thought that *p*—is done intentionally, for he does indeed provide himself with a reason for doing it and avoids thinking that *p* for that reason. However, he does not realize that his desire that not-*p* and his belief that *p* combine to motivate him to fabricate and accept that reason as warrant for avoiding the thought that *p*. If this aspect of his doings were intentional, then there would be something paradoxical about the fact that he could coherently and successfully contrive not to think what he believes but something to the contrary instead. Though motivated this is not intentional.

There is something puzzling about how one can be caused to accept reasons one would normally reject. We may not have an adequate explanation of this phenomenon, but at least we have seen that it is not peculiar to self-deception but characteristic of many sorts of behavior. Regarding the motivation of the self-deceiver, we need not suppose that there is anything (over and above the activity of deceiving oneself or the state of being self-deceived) keeping him from being aware of his belief that *p* or of his uncharacteristic violation of his own rational standards. Not being aware of these things does not mean he is prevented from being aware of them, even though if at some point he became aware of them he would either become undeceived about *p* or deceive himself about them as well. Either way we need not assume that he was prevented from being aware of them up to that point.

Self-deception is often assimilated to bad faith, and yet bad faith covers a multitude of other sins. Whereas self-deception proper concerns cognition (i.e., thinking and believing), these others involve the affects. I wish to end by proposing without argument the following parallel between self-deception and other cases of bad faith. I suggest that both consist in an uncharacteristic schism between some mental state and the corresponding occurrence. Just as the self-deceiver does not think what he believes but something that he does not believe instead, so the person in bad faith has an emotion that he does not feel and feels one that he does not have. An account of how this comes about and of just what the person does to help bring this about would not only spell out the suggested parallel between self-deception and

bad faith in general but give specific content to the idea that a person can be divided against himself.

KENT BACH.

SAN FRANCISCO STATE UNIVERSITY.