PHILOSOPHICAL PROBLEMS AND PARADOXES IN THE CONCEPT OF SELF-DECEPTION, WITH SPECIFIC REFERENCE TO PERVERSIONS OF RATIONALITY

Nelleke Bak

A Dissertation Submitted to the Faculty of Social Sciences and Humanities
Philosophy Department, University of Cape Town
for the Degree of Master of Arts

Cape Town 1987

ABSTRACT

TITLE: PHILOSOPHICAL PROBLEMS AND PARADOXES IN THE CONCEPT OF SELF-DECEPTION, WITH SPECIFIC REFERENCE TO PERVERSIONS OF RATIONALITY

CANDIDATE: BAK, Nelleke

The problem of self-deception has been described as the paradoxical state of fooling oneself into believing what one knows to be false. The epistemological paradox of believing that p and believing that not-p, the psychological paradox of intending to do what one knows one cannot do, and the ethical paradox of being both agent and victim of one's own deception arise when self-deception is based on the structure of other-deception. Traditionally the approach to these paradoxes has been either to assert that literal self-deception, as based on the structure of other-deception, is impossible and that those phenomena which we falsely call "self-deception" are merely metaphors of other-deception, or the other approach is to assert that literal self-deception, as based on the structure of other-deception, is possible with all its accompanying paradoxes. Taking as her starting point the belief that self-deception can be based on the structure of other-deception, the author aims to show that self-deception is problematic but not necessarily paradoxical and that the two traditional approaches are not necessarily exclusive. The author has placed self-deception on a sliding scale from "weak" to "hard" cases, analogous to a sliding scale of "weak" to "hard" other-deception. By means of conceptual analysis of "deception" and the comparison and evaluation of different arguments, the author attempts to explain how self-deception, as the holding of contradictory beliefs, is possible.

The thesis addresses the following questions:

1. What distinguishes self-deception from other forms of irrational belief-formation or irrational action?
2. What are the strategies employed by the self-deceiver in order to prevent his own rationality?
3. Does self-deception necessarily imply internal irrationality?
4. How is "hard" self-deception possible?
5. How is the Freudian theory of systems compatible with that of Davidson?

In answer to the first two questions the author extricates characteristics of self-deception from various related phenomena and follows David Pears' classification of three main strategies. In response to the others, she discusses the numerous theories of "weak" self-deception and in particular the theories of "hard" self-deception advanced by David Pears and Donald Davidson.

The thesis concludes that

1. Self-deception implies internal irrationality only when the agent holds a second-order principle.
2. "Hard" self-deception is possible only if the psyche is compartmentalized into mental systems.
3. The Freudian theory concentrates on the psychological aspects of self-deception, whereas the Davidsonian thesis addresses the rational aspects.

DECLARATION

I declare that this dissertation is my own, unaided work. It is being submitted for the degree Master of Arts in the University of Cape Town. It has not been submitted before for any degree or examination in any other University.

Signed by candidate

Nelleke Bak

16th day of September, 1987.

CONTENTS

ACKNOWLEDGEMENTS

My thanks are due to the University of Cape Town for a postgraduate scholarship and to the Human Sciences Research Council for financial assistance which enabled me to complete this thesis. I should also like to thank my supervisor, Dr. David Brooks, whose diplomacy and guidance have been invaluable.

INTRODUCTION

PARADOXES AND PROBLEMS OF SELF-DECEPTION

The notion of self-deception, with its air of paradox, is an odd enough idea to have elicited a great deal of attention in recent years. It is a puzzling idea, for it is not merely a non-rational phenomenon which lies outside the bounds of the rational, but it is irrational, an overthrow of reason.

In the Oxford English Dictionary "self-deception" is defined as "the action of deceiving oneself." This definition steers our understanding of the concept of self-deception in the direction of deception of others, and tempts us to base our explanation of self-deception on the model of other-deception. If other-deception then is getting someone else to believe something one knows is not true, then self-deception is getting oneself to believe something one knows is not true. Therefore, if self-deception is "the action of deceiving oneself", then we have a case in which one and the same person deceives himself through one and the same action (i.e. the action of self-deceiving) with regard to his own beliefs. The one person is at the same time both deceiver and deceived, agent and victim, blameworthy and blameless, aware and ignorant by persuading himself to believe something he knows to be false. Raphael Demos rekindled the recent interest in the philosophical investigation of the concept of self-deception by describing the phenomenon in the following way:

> "Self-deception exists, I will say, when a person lies to himself, that is to say, persuades himself to believe what he knows is not so. In short, self-deception entails that B believes both p and not-p at the same time. Thus self-deception involves an inner conflict, perhaps the existence of a contradiction." (Demos, 1960, p.588)

It would seem then that self-deception for a rational, sane person is not possible, that "self-deception" contradicts itself and that self-deceivers cannot literally deceive themselves. The fact that the word "self-deception" occurs in our vocabulary and the fact that we use the word frequently in our everyday language is in no way a guarantee of the actual existence of what the word literally means, i.e. the act of consciously getting oneself to believe something which one at the same time knows to be false. It has been argued that such a phenomenon, taken literally, is impossible (1), and yet we do use the word "self-deception" often and come across cases of self-deception in everyday life. The following are examples of what, I think, people would call typical cases of self-deception: the wife who, in the teeth of the evidence, will not admit that her husband is being unfaithful; the scientist who wilfully ignores telling evidence which will refute his favoured hypothesis; the athlete who is unable to reconcile himself to his deteriorating performance and waning powers; the person with cancer who refuses to acknowledge the fact that he needs medical attention, the physicist who denies that his research in developing nuclear arms implicates him in any way in promoting warfare; the girl who has been spurned by her lover, and then pretends that she never wanted to marry him anyway. But to present a list of typical examples of self-deception is in no way to offer an explanation of what these cases are typical of and fails to show any understanding of how self-deception is at all possible.

In this chapter I shall, by way of an introduction, note the different meanings of the word "deceive", before going on, at a later stage, to look at the implications in more detail. The different interpretations of "deceive" have given rise to different theories with different emphases. By briefly mentioning some of the theories, I aim to show that self-deception

has no clear-cut boundaries and it is this which leads to divergent descriptions and analyses. Before launching into an investigation of self-deception, it is necessary to define the meaning of the word "deception" clearly. It is important to distinguish between the literal meaning of prefixing "self" to "deception" and the everyday, common meaning of "self-deception". A little later on in this chapter, I shall look in more detail at the denotation and connotation of "deception" and thus "self-deception". I have already mentioned the definition of "deceive" as getting someone else to believe something one knows to be false, but the Compact Edition of the Oxford English Dictionary offers a second definition of the word "deceive", viz. "to cause to believe what is false". These two meanings of the word "deceive" lead to two distinct meanings of other-deception. First of all, there is deception 1 which rests on the meaning of deceive as "to cause to believe what is false". In terms of other-deception then, B is deceived if B believes falsely that p because of something A has done. In other words, A has caused B to form and hold a false belief. For example, A accidentally misreads something from the newspaper and B, therefore, comes to hold a false belief; or A mishears a snippit of information, tells B and B as a result forms a false belief; or A mumbles indistinctly about something which causes B to form an erroneous conclusion. My reason for noting these examples is to show that A can deceive B without intending to do so. In fact, A is unaware that B has been deceived because of A's accidental misreading, or A's telling B something which is false but which A believes to be true, or A's indistinct mumblings. Then there is deception 2 which is based on the meaning of deceive as an intentional action. In terms of other-deception then, B has the false belief that p because A who has the true belief that not-p wants B to believe falsely that p and gets B to believe falsely that p. A's desire to knowingly deceive B makes this a case of intentional deception.

So, referring to deception 1, "B is deceived about p" may mean simply that B has been caused by someone or something to be in error with respect to p. Whoever did the deceiving, need not have done so intentionally. If B is mistaken about p because of something A has said or done (or has failed to say or do), it does not necessarily imply that A purposely cultivated a false belief in B which A knows to be false. Even though A has been the cause of B's holding a false belief, A need not necessarily have done this intentionally. This means that not all deception entails intentional deception. Intentional deception is, of course, the central case of deception, but I want to stipulate the different meanings and uses of "deception". If "deception" and, therefore, "self-deception" are not clearly defined, the confusion could lead to many problems.

Basing one's understanding of self-deception on the model of other-deception then involves having two different meanings of self-deception as well. Self-deception 1 based on the model of deception 1 thus means that "A is self-deceived" when A believes falsely that p because of something A has done. A is, therefore, in error with respect to p. Self-deception 2 based on the model of deception 2 is the more interesting and paradoxical case in which "A is self-deceived" when A believes falsely that p because, believing that not-p, he wants to believe that p and gets himself to believe that p. So "self-deception" can mean to deceive oneself (self-deception 2), as well as causing oneself to be deceived about oneself, to be in a state of error about one's own thoughts and feelings (self-deception 1).

Hamlyn notes that the double-life of the meaning of "deception" can lead to confusion about self-deception.

> " .... self-deception, as we typically understand it, cannot be construed as the parallel of those forms of deception in which we deceive someone by, for example, giving him by mistake erroneous information. Hence, it may seem reasonable to attempt to construe self-deception on the pattern of deception of others when this involves, as we might put it, being false to them."
> (Hamlyn, 1971 a, p.47)

Clearly, what is of interest in a philosophical investigation of self-deception is the case of "knowingly being false", of intentionally deceiving oneself. However, before looking in more detail at unintentional and intentional deception, it is necessary to define the meaning of "deception" a little further.

Another confusion which arises is that between "deception" and "deceit". The word "deception" immediately conjures up an association with "deceit", but the association is not always a necessary one. Not everything that deceives is deceitful. The word "deceit" has a moral implication in that the deceitful agent wilfully causes someone else to accept a falsehood, whereas the word "deception" can in some cases be free from any moral condemnation, but may merely point to a state of error. Not everyone or everything that brings about a state of error (e.g. mistaken perception) does so with moral (or rather, immoral) intentions. In a desert a mirage may deceive the thirsty traveller into thinking that an oasis is nearby, or the Müller-Lyer visual illusion with its arrangement of two lines of equal length may deceive the observer into believing that one line is longer than the other. Even though the deceptive arrangement causes the innocent eye into believing the equal lines to be of unequal length, we would not call the deceptive Müller-Lyer visual illusion deceitful. We may also talk about the deceptively slow ball, but there is no moral implication in such an observation. (Champlin, 1979, p.87). So, the self-deceiver is

not always a cunningly deceitful liar, but he can be someone who sets up a deceptive screen between himself and a too painful reality. The screen hides or distorts the unwelcome truth and the agent himself is then taken in or deceived by this distorted, deceptive appearance of his own making or he might form and hold on to a favoured belief, despite evidence for a counter-belief. But this does not really distinguish a case of self-deception from a bit of harmless wishful thinking or a case of mere oversight. In Chapter 2 I shall draw the distinction more clearly, but it does seem that self-deception proper seems to imply a moral element of dishonesty with oneself. Self-deception proper then is not merely being in a state of error, but implies an intentional cultivation of a false belief, a deliberate lying to oneself. Although this case of self-deception, with its moral implications, forms the most fascinating and most central case of self-deception, it is not representative of all cases of self-deception, which are not all coloured by moral tones.

It is not surprising that such slippery terms as "self-deception" and "self-deceivers" have been described by a colourful collection of ideas and images. Self-deception has been compared to the act of concealment, self-deceivers hide or bury or camouflage the truth from themselves.(2) Visual images include descriptions such as self-deceivers being blind to the truth, preventing themselves from seeing the facts, intentionally obscuring their own view of reality. Other descriptions appeal to auditory comparisons, such as self-deceivers refusing to listen to or hear the truth, or being deaf to what they do not want to hear. Using recognition terms, philosophers describe them as avoiding the unpleasant belief, refusing to identify or recognize what they suspect.(3) In the language of the theatre they are described as playing two conflicting roles, as acting out their irrational beliefs; they are accused of pretending that

some strongly desired thing is so as to increase the illusory realism of their theatrical project.(4) In the language of rationality, self-deceivers are seen as refusing to make explicit in language, a "spelling-out", of some feature of the world, or of persuading themselves of something which is not, or of fooling themselves.(5) Affective terms describe self-deceivers as being blinded by passion or misled by desire. In the language of consciousness, it is said that they ignore the harsh reality, or suppress unwelcome ideas and topics. Self-deceivers are described in terms which invite evaluative judgments, such as lying to themselves, as being guilty of dishonest role playing, inauthenticity or, indeed, as practising self-deceit.(6) In social terms we see self-deceivers as being in a certain engagement in the world, a refusal to avow or confess something to themselves, as desiring to appear existentially successful.(7)

Following the diversity of descriptions of self-deception are a number of divergent theories on what self-deception is, how it is at all possible, and why it takes place. However, there is a common starting point or basic description of self-deception upon which various theories are built. The common assumption, anyhow in clear-cut cases of self-deception, is that to deceive oneself entails persuading oneself to hold a belief believed to be false.(8) The various problems coupled to this assumption are to explain just <u>when</u> false beliefs constitute a case of self-deception rather than, say, wishful thinking and <u>how</u> self-deceivers form their false beliefs (the initiation of the project of self-deception) and how they <u>maintain</u> their false beliefs (the perpetuation of the state of self-deception). With such a shifty topic as self-deception, it is not surprising that theories have developed into various directions, stressing different aspects of the phenomenon. One theory sees self-deception primarily as a function

of what one thinks (or avoids thinking about) as opposed to what one believes. Self-deceivers do not necessarily hold conflicting beliefs but rather avoid the unwelcome thought(9); they rarely have the intention of deceiving themselves - they are motivated by desire to manipulate data(10); they neither just believe, nor both believe and disbelieve, nor just disbelieve what they present themselves as believing - they are genuinely uncertain about the evidence and formulate the favoured belief not against the evidence, but within the latitude which the inconclusive evidence allows(11); they are involved in some kind of "conflict state", but they need not be aware of it(12); they form their beliefs in "belief-adverse circumstances", circumstances in which the evidence is heavily weighted against the unwelcome belief(13); the self-deceivers do not necessarily, jointly and simultaneously, fulfill the conditions for full belief and for full disbelief, there is only partial satisfaction of the different criteria for belief and for disbelief(14); they are in a conflict state, but the wish for the favoured belief "screens" the unwelcome belief from consciousness(15); they both know and do not know, believe and do not believe(16); self-deception is possible because there is a causal mechanism involved whereby repetition of statements which are false (and known to be false) eventually produces a belief in their truth(17); self-deception involves two conflicting beliefs but both can survive because they are kept apart by their failure to interact rationally with each other(18); self-deception is not a matter of knowledge and belief - it is rather a refusal to "spell out" a specific engagement in the world(19); self-deception is reinterpreted as attempts at self-modification taking place within a "zone of indeterminancy" which allows the self-deceiver to pick out selective evidence(20); self-deception cannot be studied in isolation of the motivational context in which it occurs(21); the significance of motives in self-deception has been overvalued - it is the

deception, and not the motives for it, that is the essence of self-deception(22); self-deception occurs when the individual's self-interpretation disagrees with the community's standard view of that individual(23); self-deception is never more than a metaphor, not a real description, but a licence to improvise(24).

Although valuable insights have been yielded by all these diverse theories, many of the theories emphasize only a few cases of self-deception (e.g. desire-motivated cases or "weak" cases) while ignoring or neglecting others (e.g. intellectual distortion of the evidence or cases in which the false belief is formed in the teeth of conclusive evidence for the unwelcome but true belief). Or a single aspect of self-deception (e.g. the intentionality of self-deception) is singled out while other elements (e.g. the mechanisms of self-deception) are not discussed. The long list of the various examples of self-deception, the different terms in which it is described and the divergent theories of self-deception illustrate the fact that self-deception can take innumerable forms, either in self-deceptive behaviour or self-deceptive belief formation. The different meanings and different uses of "deception" and "self-deception" in everyday language further compound the problem. Furthermore, "self-deception" can be applied to either the process of initiating the self-deceptive project or can refer to the state of being in self-deception. The boundaries of self-deception are vague, with no clear-cut distinctions always drawn between wishful thinking, akrasia or even unintentional ignorance or inadvertently biased beliefs.

And even when the self-deceiver has persuaded himself to hold a false belief, the way in which he persuades himself can take many different forms, the belief can be held in a variety of different manners, and the context within

which the belief is either formulated, held or acted out can change continuously. Moreover, the motives which give rise to self-deception can vary greatly from ambition, greed, fear, conscience to general desire. Added to this is the fact that self-deception is not necessarily a case of beliefs only - it can amount to wilful ignorance, avoidance of certain evidence, emotional detachment, or even self-hypocricy (not really having a belief, but pretending to have one).

With such a motley collection of approaches, emphases, perspectives and meanings, it will be necessary to stipulate exactly what the boundaries of my investigation will be, the meaning of the word "self-deception" and from which perspective I shall view the problem of self-deception.

First of all, I shall base my notion of self-deception on the model of other-deception.

> "If deceiving someone else is getting him to believe something one knows is not true, then deceiving oneself would appear to require getting oneself to believe something one knows is not true." (Mele, 1982, p.159)

However, it is important to note that "when one man deceives another, it need not be the case that he intends to deceive him."(Champlin, 1977, p.292) Champlin cites the case of a spy who, without his knowledge, is given false information about secret plans for an invasion. His deceivers, the heads of Special Operations, are gambling on the spy's being captured by enemy forces, cracking under torture, and leaking the plans of the alleged invasion to the enemy. So, the Special Operations Chief deceived the spy into believing that he had genuine information; the spy (who did duly crack under torture) deceived the enemy into believing his information to be genuine. Although we can say that the Special Operations Chief intentionally

deceived the spy, we would be wrong to say that the spy intentionally deceived his tortures, even though he did deceive them. The spy deceived the enemy without knowing that he was doing so.

If the second definition of "deceive", as found in the Compact Edition of the Oxford English Dictionary, is employed (i.e. "to cause to believe what is false ...") then A may induce a false belief that p in B, and thus deceive him in this sense, without A's knowing or even believing that p is false. Indeed, A may believe that p is true, and he may have intended to communicate this to B by telling him that p. Although this is not a case of intentional deception, it is a case of deception nonetheless. There are other examples of unintentional other-deception. A may unintentionally neglect to inform B of a change, thereby causing B to form a false belief. Unintentional other-deception may also be due to the inadvertent use of an ambiguous word or gesture, forgetfulness, misperception, indiscriminate haste or plain bad luck.

Although there are cases of unintentional other-deception and, therefore, of unintentional self-deception, since I shall be basing my notion of self-deception on the model of other-deception (see Chapter 4 for a discussion of unintentional self-deception), the cases of self-deception and other-deception I am really interested in are those of intentional deception.

> "The suggestion that self-deception should be modeled after unintentional interpersonal deception runs the risk of oversimplifying matters. Unintentional interpersonal deception may be quite accidental. But self-deception seems to be motivated by desires or fears of the agent-patient." (Mele, 1983, p.367)

If, for example, Tim unintentionally misreads the results of the horse races, thereby causing Ed to have the false belief that he has won a lot of money,

then Tim caused Ed to be mistaken or deceived about the matter and is said to have deceived Ed (albeit unintentionally) about it. Similarly, if Tim were to misread the results due to unmotivated carelessness, thereby causing himself to have a false belief and, hence, to be deceived, he is unintentionally self-deceived. But intentional self-deception seems to entail a desiring element, something which occurs because the agent wants it to be the case that p. So, if Tim, who has been having a run of bad luck on the horses and who desperately wants to win some money, misreads the results as being in his favour due to his desire for the money, we come much closer to a case of intentional self-deception.

Cases of unintentional other-deception employ a "loose" sense of "deception" (i.e. "to cause to believe what is false ..."), for one can cause without wanting to cause or even without knowing that one will cause or has caused someone else to hold a false belief. Intentional other-deception, on the other hand, appeals to a stricter sense of "deception", that of wanting to deceive the other person. The deceiver knowingly embarks on a belief-misleading project, "an attempt to manoeuvre some intended victim into taking what is only an appearance for a reality, so that the victim forms a belief leading him to act (or think) in some way that is desired by the deceiver but that he would not act in if he knew the reality." (Kipp, 1980, p.313).

However, intention needs more than mere desire. Davidson distinguishes the various ways in which intention operates. A may have the desire to deceive B, but through an unrelated action of A's (one not calculated to deceive), B becomes deceived. In this case, A did not deceive intentionally. However, in the second case, A has the desire to deceive B, believes that action x will cause B to become deceived, and deliberately performs action x

which leads to the successful deception of B. In this case, A has deceived B intentionally. Then there is the third case in which A has the desire to deceive B, believes that action x will cause B to become deceived, but another inadvertent action by A causes x to come about which still leads to the successful deception of B. However, in this third case we would not accuse A of intentionally deceiving B because A's desire and belief did not cause the action x in the right way (as they did in case 2). Furthermore, for a case of intentional other-deception, knowledge is a prerequisite. The degree of knowledge necessary for a case of intentional other-deception, however, varies from suspicion in "weak" cases of deception to absolute certainty in "hard" cases. Lastly, deception includes the notion of effect. Deception implies the success of inducing an erroneous belief in the other's mind.

Haight offers the following description of intentional other-deception:

> "... (I)f A deceives B, then for some proposition(s) p, A knows that p; and either A keeps or helps it to keep B from knowing that p, or A makes or helps to make B believe that not-p, or both." (1980, p.9)

I shall use this as a general definition of other-deception and shall look at various ways in which the deceiver prevents the deceived from knowing the truth, or how the deceiver distorts or manipulates the truth, or how he does both. I shall use the above definition, as well as the following ways in which one person deceives another, as a model for self-deception. Just how the analogous ways of other-deception manifest themselves in self-deception will be discussed in Chapter 4 which will concentrate on "weak" cases and Chapters 5 and 6 which will look at "hard" cases.

The following ways in which one person can intentionally deceive another are arranged in an order of increasing severity, i.e. from five "weak" cases

to a "hard" case.

1. A withholds evidence from B to cause B to have a false belief. For example, Tim deliberately does not tell Ed that a mutual friend has been acquitted on a charge of fraud, thus causing Ed to hold the belief which A wants him to hold, i.e. that the friend is guilty.

2. A selectively exposes B to or steers B away from certain telling evidence, thus causing B to have the false belief which A wants him to have (or A wants to prevent B from holding the true belief).

3. A emphasizes certain evidence to B causing B to have the false belief A wants him to have.

4. A generates pseudo-evidence to B causing B to have the false belief A wants him to have.

5. A distorts available evidence to B causing B to have the false belief A wants him to have.

6. A directly tells a falsehood, lies, to B causing B to have the false belief A wants him to have.

Cases 1 to 5 are "weak" cases of other-deception, in that there is a certain amount of "leeway" or "latitude" in which the manipulated evidence can be interpreted. If, for example, A wants B to think that his wife is being unfaithful, A may deliberately steer B into a restaurant where B's wife is having lunch with a male friend. B's noticing his wife is no guarantee for B's forming the false belief of his wife's infidelity. He <u>may</u>, of course, form the false belief but, on the other hand, he may interpret the lunch as an official business appointment. A gambles on the <u>chance</u> that B will form the false belief but he has no certainty of this happening.

Case 6, however, is regarded as a "hard" case of other-deception in that it is not a case of manipulating or exploiting available evidence, but of fabricating false evidence which is then directly presented to the victim as being true evidence, i.e. A would declare to B, "Your wife is unfaithful", even though A would know that this is not the case.

Other-deception is, therefore, a relation between two persons, one who knows the truth and another who is ignorant of it, or one who believes what the other knows is false. The following is a summary of the relation which holds between deceiver and deceived:

1. A withholds/manipulates/distorts evidence or makes a statement (p) to B which A knows or believes or suspects to be false or A withholds from making a statement (not-p) to B which A knows or believes or suspects to be true.

2. A intends B to believe that p.

3. B comes to believe that p as a result of A's behaviour in the manner A intended.

Therefore, for successful intentional other-deception to take place, the above conditions will apply. The conditions are by no means undisputable, but merely serve as a general analysis of other-deception(25) and thus as a general model of self-deception.

If the notion of self-deception is to be modelled on the structure of other-deception, it seems as though it runs into some serious paradoxes. Before an answer can be found to these paradoxes, it will be necessary to look at what exactly the paradoxes are. In this chapter I shall confine the

discussion to an elucidation of the problems themselves, without attempting to find an answer to them at this stage. I want to base my notion of self-deception on that of other-deception, but how can conditions (1) to (3) all be satisfied in the case of one person? How can one believe a statement is true when one at the same time knows or believes it to be false? (conditions (1) and (3)). Or how can one intend to do something which one knows one cannot do? (condition (2)). The symmetry between other-deception and self-deception can be expressed in the following questions: "Can you deceive someone else into believing what he already knows to be false?" and "Can you deceive yourself into believing what you already know to be false?" We can answer a fairly confident "yes" to the first question (we need look only at the success of propaganda and brain-washing), but hesitate to answer the same to the second question, because of the paradoxical implications. I shall classify the various paradoxes into three main groups: the epistemological paradox which entails that the self-deceiver both believes p and believes not-p; the psychological paradox in which the self-deceiver cannot be aware of what he is doing (if the deception is to be successful), nor can he be unaware; and the ethical paradox which implies that the self-deceiver is both sincere in the profession of his belief and insincere in employing various devices to conceal the truth.

First of all, the epistemological paradoxes centre on knowledge and belief. Is it possible to persuade oneself intentionally to believe what one simultaneously knows to be false? To act on such an intention requires the self-deceiver to use his very grasp of the truth in order to negate that very same truth. Surely if the self-deceiver knowingly embarks on a subversive attempt to undermine the truth his project is doomed from the start?

But there are examples of successful self-deception. What does the success of such a paradoxical project imply? Can we wilfully distort our beliefs even when there is counter-evidence for those beliefs? Can we simultaneously and knowingly hold directly contradictory beliefs or do we have two separate minds, one doing the deceiving and the other being deceived? When self-deception occurs it seems as though one "self" knows something, but prevents the other "self" from knowing it, or even getting the other "self" to believe the opposite of what the first "self" knows. The main problem of the epistemological paradox can be expressed as follows: "Could a self-deceiver ... bring to consciousness two directly contradictory beliefs at the same time, formulate them in words, be clear about their meaning, believe they are contradictory, be sane and not intoxicated, and still hold them?" (Martin, 1986, p.23)

The epistemological paradox is the one which has attracted the bulk of philosophical interest. It has been expressed in many ways from Demos's "self-deception entails that B believes both p and not-p at the same time" (1960, p.588) to Sartre's formulation of self-deception as "bad faith". Sartre interprets other-deception in terms of lying to another (case 6) and, basing self-deception on the model of other-deception, he interprets self-deception as "a lie to oneself". Only what changes everything is that in self-deception "it is from myself that I am hiding the truth" and adds that "the essence of the lie implies in fact that the liar actually is in complete possession of the truth which he is hiding. A man does not lie about what he is ignorant of". (Sartre, 1958, p.48) To interpret the phrase "self-deception" literally then, would require us to regard the self-deceiver as someone who knows that something is not so, and yet persuades himself that it is so, just like in other-deception, except that the self-deceiver is both agent and victim. From here it is a short step to saying that this

implies that the self-deceiver must believe the conjunction of the two contradictory propositions. Since this is impossible for any sane person, it is argued by those who deny that literal self-deception can exist, that self-deception is impossible. If we follow the arguments which deny the possibility of literal self-deception then there are two assumptions: first of all, it is necessary for literal self-deception that the original belief should exist right up to the end of the project of self-deception, which entails not only the formation of the rational belief and its opposite, but also the safeguarding of this belief against telling evidence. Secondly, someone who believes two contradictory propositions believes their conjunction. Both these assumptions should not be granted so easily. Later on I shall explain how the self-deceiver, with the passage of time, may come to forget the initial belief (Chapter 5), and how the self-deceiver can still be self-deceived, without his putting the two contradictory beliefs together. (Chapters 4 to 6)

The alleged contradiction(26) of p and not-p can be expressed in a number of different forms:

(1) aBp+-aBp where it means that Andrew believes p and it is not the case that Andrew believes p. This is an outright contradiction. Since this is an obvious impossibility, this formulation of self-deception is ruled out.

(2) aBp+aB-p where it means that Andrew believes p and Andrew believes not-p at the same time. Although aBp+aB-p seems like a contradiction, it is not necessarily so. Priest points out that "Many, in fact most, of us believe contradictions. The person who has consistent beliefs is rare. If someone has never found that their beliefs were inconsistent,

this probably means that they just have not thought about them long enough." (1986, p.102) It is, in fact, this formulation of self-deception on which I have based my study and in Chapters 4 to 6 I shall argue for the possibility of the self-deceiver believing p and believing not-p. Logic prevents both beliefs from being true at the same time, but it does not prevent both beliefs from being held. I shall show, however, that not all cases of inconsistent beliefs are cases of self-deception.

(3) aBp+a-Bp where it means that Andrew believes p and he disbelieves p. It would seem that logic would rule out the possibility of two contrary attitudes—pro and con, believing and disbelieving—existing in the same person. However, to believe and to disbelieve does not necessarily point to a contradiction. Andrew may have heard good arguments both for and against capital punishment and is swayed to believe p and to disbelieve p, depending on which speaker is arguing.

(4) aB(p+-p) where it means that Andrew believes that a contradiction is true. This is the limiting case of the epistemological problem, and the one on which philosophers concentrate in their arguments for the impossibility of literal self-deception. This form requires the self-deceiver not only to hold two conflicting beliefs but to conjoin them as well. Audi notes that "While it seems possible for one to have beliefs of incompatible propositions, it is not clearly possible for one to believe two propositions one <u>believes</u> are incompatible." (1982, p.138)

> In other words, a conflict of beliefs seems possible, but a belief in what is conflicting seems not.(27) However, I shall argue that it is not necessary for the self-deceiver to conjoin the two conflicting beliefs in order for his project to qualify as a case of self-deception. It is obvious that many people have inconsistent beliefs, p and q, but to believe in a contradiction, i.e. (p+-p), is quite different to holding incompatible or inconsistent beliefs.

A further complication arises. Case (2), aBp+aB-p, can be expressed in the following two statements:

(1) Andrew believes that he has red hair. (p)

(2) Andrew believes that he does not have red hair. (not-p)

The problem is not just two apparently contradictory statements, (Andrew believes both proposition p and its negation not-p) which he holds without putting them together, but he believes not-p because he believes p. If Andrew did not have the belief that he has red hair, he would have no need to deceive himself about it. "Self-deception is notoriously troublesome, since in some of its manifestations it seems to require us not only to say that someone believes both a certain proposition and its negation, but also to hold that the one belief sustains the other". (Davidson, 1985, p.138) In Chapters 5 and 6 I shall explain how someone can have beliefs like in (1) and (2) without his putting p and not-p together, even though he believes not-p because he believes p.

The second paradox of self-deception, the psychological paradox, encompasses most of the problems posed in the epistemological paradox, with the added

complication of the intention with which the self-deceiver embarks on the process of self-deception and sustains the state of self-deception. Other-deceivers act on the basis of what they know in order to hide or distort this knowledge to their victims. But then the self-deceiver is required to use his knowledge or suspicions of some truth as a basis for intentionally ignoring that truth. "It seems especially puzzling that some self-deceivers can systematically ignore what is so easy for them to grasp." (Martin, 1986, p.24) Pears observes in Motivated Irrationality that what makes self-deception seem paradoxical is not merely what the self-deceiver believes or the irrationality of his contradictory beliefs or his belief unfounded on evidence, but also what the self-deceiver does which, according to Pears, is intentionally cultivating an irrational belief. The self-deceiver intentionally accomplishes the feat of his holding contradictory beliefs. He designs a plan which will successfully conceal the unwelcome belief from him, and yet to formulate the plan he must be aware of his contradictory beliefs and the irrationality of his favoured belief, which brings us back to the epistemological paradox. He knowingly forms the self-deceptive plan and yet, perplexingly, for the plan to be effective he cannot be aware of it. The paradox of motivated irrationality is particularly evident when the action is intentional and free. The self-deceiver does not appeal to others to do the job for him, such as hypnotists or brain surgeons, but he does the job himself. The paradox would disappear if the subject were hypnotized or brainwashed. Even if he intentionally went to a hypnotist and instructed him to "implant" a false belief, we would not have a case of self-deception, for an external factor has done the "implanting" of a false belief. The question may arise whether the paradox of self-deception is preserved if the agent intentionally practises hypnosis on himself. Self-hypnosis used for successful self-deception does not contain a paradox because the unwanted belief is expunged

and does not sustain its opposite. The paradox would also disappear if the self-deceiver fails to recognize the irrationality of his beliefs, due to fatigue, negligence, incompetence, impetuosity or plain bad luck. "Conscious irrationality is paradoxical only if it is avoidable. There is nothing suprising about a consciously irrational belief that is truly obsessional." (Pears, 1982b,p.162) If a person were not competent to detect his errors, or to diagnose the irrationality of them, the paradox of self-deception would disappear. The self-deceiver who poses a problem is the one who is intentional, free (from external forces) and competent to detect his own irrationality. How then can a rationally competent person freely and intentionally form a belief against the total weight of the evidence available to him? How can a rationally competent person freely and intentionally lie to himself? Sartre expresses the paradox as follows: "The liar intends to deceive and he does not seek to hide this intention from himself nor to disguise the translucency of consciousness" and concludes that "the liar must make the project of the lie in entire clarity and that he must possess a complete comprehension of the lie and of the truth which he is altering." (Sartre, 1958, p.48-49) In forming and trying to act on the intention to ignore/conceal/distort would not using the knowledge of what is to be concealed/distorted block the very effort of concealment or distortion? In Sartre's words, deceivers apparently "must know the truth very exactly in <u>order</u> to conceal it more carefully". (1958, p.49) Moreover, the self-deceiver has knowledge of the intention to deceive but would not this self-knowledge thwart the project of deception? Sartre, who refers to deception as "bad faith", poses the problem as "that which affects itself with bad faith must be conscious (of) its bad faith since the being of consciousness is consciousness of being [what it is]." (p.49).

The third paradox of self-deception, the ethical paradox, concentrates on the problematic moral status of the self-deceiver. As active agents, self-deceivers seem guilty of practising deceit and are guilty of any harmful consequences of that deceit; as passive victims, self-deceivers seem to be innocent victims. When the agent deceives himself, we have a case where one and the same person, with regard to one and the same action of his, is both morally blameworthy (in his role as deceiver) and deserving of moral sympathy (in his role as a victim of deception). "Viewed as liars, they appear insincere and dishonest; viewed as victims of a lie, they appear sincere and honestly mistaken. As deceivers, they seem responsible and blameworthy for cowardly hypocrisy; as deceived, they apparently deserve compassion and help in gaining full awareness of the guile perpetrated on them." (Martin, 1986, p.29)

Gardiner discusses (1970, p.221-3) Butler, who presented one of the earliest accounts of the moral paradox of self-deception. Bishop Butler, an eighteenth century cleric, looked at self-deception as "self-deceit" in a predominantly moral context. He was concerned with cases in which an individual might be said not to recognize such things as his own faults and failings. Butler concludes that in self-deception, being a species of dishonesty and "falseness of heart", the deceiver is more guilty than innocent. Possibly because of his clerical preoccupations, Butler regards self-deception almost entirely as a case of wrongful action and the application of double standards. Self-deception is seen as involving some sort of conflict between good and evil forces, the force of conscientious scruple which is defeated by the force of selfish desire. The moral paradox is not so easily dismissed, for self-deception displays itself in situations in which the moral issues are often shadowy and ill-defined, where it is hard to draw a sharp line indicating the point at which right ends and wrong

begins. Moreover, self-deception is not always a triumph of "selfish desire", e.g. a mother who is petrified of water could deceive herself into thinking that she is not scared of plunging into the river and so save her child from drowning.

The epistemological, psychological and ethical paradoxes cover the three main problematic areas of self-deception. I shall, however, mention a few further problems involved not only in self-deception itself but also in the method of approach to and the description of the concept of self-deception.

The epistemological paradox of self-deception focuses on those agents who are interested in the truth (albeit an interest in distorting or avoiding the truth). But people, in general, are not always interested in truth. Some theories of self-deception seem to be inseparably dependent on the assumption that the self-deceiving agent has a compelling interest in the truth of what he appears to believe. Gardiner (1970, p.241) points out that human attitudes and behaviour are often treated as being more purposeful, more reflective, more deliberate, more under control of the conscious will, than experience suggests them to be the case. One way of solving the problem of self-deception is to interpret the intention of the self-deceiving agent and his knowledge of the falsity of the favoured belief <u>not</u> as the sort of fully-fledged conscious deliberations that cause trouble and give rise to the paradoxes. In most cases of self-deception, Gardiner states, there is often no apparent realization on the subject's part, or at best only an intermittent one, that he is falsifying things or making them look other than they are. People are naturally open to simultaneously holding opposing beliefs. Most people hold inconsistent beliefs and most people do so with relative ease. "Who but a Descartes would think possible the task of sorting out all of his or her beliefs? We unabashedly admit that we do not

know the whole range and all the implications of our beliefs; and we are not all of us good enough logicians to be sure that we would recognize inconsistencies among the beliefs of which we are aware". (Hanson, 1986, p.109) If self-deception is an attempt to mask the irrationality of thought processes, self-deception then is practised by agents who are particularly keen on evaluating themselves as rational beings. Self-deception is, therefore, a strategy used only by rational and consistency-loving people who want to disguise the fact that they are falsifying things or holding incompatible beliefs. Paradoxically then, self-deception is seemingly a vice peculiar to rational beings, while the irrational majority are free from its temptations. If one orients one's beliefs on reality, one is rational, but to orient one's beliefs solely on one's desires is to make oneself susceptible to self-deception. Therefore, Kipp concludes (1980, p.312) desiring to believe oneself consistent when one knows one is not is, by definition, incompatible with really being rational and consistency-loving. Those who are quite ready to tolerate recognized irrationality and inconsistency in their beliefs are seemingly not those who want to disguise the existence of their irrationality by means of a self-deceptive project; but those who strive to be rational, consistency-loving beings are seemingly the ones who embark on an irrational project of self-deception in order to mask the existence of their irrationality, not only of their holding two opposing beliefs but also of their project to deceive themselves about the belief that they hold two opposing beliefs. To resolve the paradox of literal self-deception, it is necessary to show that a person might hold two opposing beliefs as well as to show how one of the two opposing beliefs (the favoured belief) is the <u>result</u> of a self-deceptive project. Self-deception is not merely a strategy for reconciling simultaneously held, conflicting beliefs, it is also strategy for deceiving oneself about one's belief that one holds conflicting beliefs. Self-deception, therefore, seems to require <u>primary</u>

deception about some unwelcome belief and also secondary deception about the resultant unwelcome belief that one now holds two incompatible beliefs. This could lead to the danger of an infinite regress of deceptions.

A further problem in the method of approach in the examination of self-deception is our inability to look directly into another person's thoughts. We can merely deduce from his words and actions what his thoughts are, and assume that he adheres to the same principles of rationality as we do. We cannot empirically observe and measure someone's thought processes and their contents directly. We can observe only his speech and behaviour. In order to explain and predict the behaviour, verbal and otherwise, of other people, we attribute beliefs, purposes, motives and desires to them and describe these in the light of the most unified and intelligible scheme we can contrive. Even speech does not yield direct access into a person's belief-system for speech itself must be interpreted by both the speaker as well as the listener. But the multiplicity of mental factors that produce behaviour and speech makes interpretation extremely complex and difficult. The solution to this problem is "to assume that the person to be understood is much like ourselves... We start out assuming that others have, in the basic and largest matters, beliefs and values similar to ours... (U)nless we can interpret others as sharing a vast amount of what makes up our common sense we will not be able to identify any of their beliefs and desires and intentions, any of their propositional attitudes." (Davidson, 1982, p.302)

Our ways of describing, understanding and explaining psychological states and events give rise to paradoxes of irrationality. In order to solve the paradoxes of self-deception, we can turn to the Theory of Mental Systems (see Chapters 5 and 6) which postulates two semi-autonomous systems of the mind with a causal relation, but between which the logical relation has

failed. Although Freud looked for terms from technology and mechanics to apply to mental events we must ask just how far the workings of the mind can be explained by strict, deterministic laws (such as hold in the worlds of technology and mechanics) as long as the mental events are described in mental terms. Davidson points out that the realm of the mental cannot form a closed system(28). If we enter the mental realm with only the causal relation holding between mental events, and to a certain extent ignore the logical relations between the descriptions of those mental events, "we enter a realm without a unified and coherent set of constitutive principles: the concepts employed must be treated as mixed, owing allegiance partly to their connections with the world of non-mental forces, and partly to their character as mental." (Davidson, 1982, p.301)

Furthermore, the Theory of Mental Systems allows inconsistent or conflicting thoughts, beliefs and desires to exist in the same mind, while basic methodology aims at an interpretation that is consistent and intelligible. It is not too difficult to explain small deviations from reality, but when we are faced with serious digressions from reality and consistency, we have great difficulty in trying to describe and explain what is going on in mental terms. The problem of methodology encountered in the study of self-deception can be summed up as follows:

> "The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all." (Davidson, 1982, p.303)

In other words, if an adequate analysis of self-deception is to be offered, it must steer a very delicately balanced path between rationality and irrationality.

The question I am going to turn to now is whether the above paradoxes are paradoxes in the philosophical sense, contradictions that cannot describe possible situations, or whether they are paradoxes in the literary sense, merely seemingly absurd statements that actually turn out to be coherent and to express genuine possibilities. The paradoxes, thus, point to a two-fold consideration: On the one hand, the paradoxes portray self-deception in the literal sense as centring on conflicting beliefs and intentional irrational belief-formation and, on the other hand, the agent's evasion of acknowledging to himself what he suspects or knows to be true. The former consideration seems to point to the impossibility of literal self-deception, whereas the latter consideration points to manifestations of this phenomenon in everyday life. Attributions of self-deception are common enough, but it does seem curious that people are so comfortable with this paradox and that they are not at all reluctant to employ a concept which seems so obviously incoherent. The paradox of self-deception can be solved by either denying that literal self-deception is possible or by explaining it away by elaborate means, thus interpreting the term "self-deception" as a metaphor. According to the literal interpretation of self-deception, the agent must believe in the conjunction of two contradictory propositions since this interpretation of self-deception implies that the self-deceiver both believes that something is so, and <u>at the same time</u> persuades himself into believing that it is not so. Our rational faculty prevents us from believing the conjunction of two contradictory propositions and since the self-deceiver must believe in this, it is argued, literal self-deception is not possible. However, as I have noted before, I shall argue for the possibility of genuine self-deception, without claiming that the belief in inconsistency (as opposed to inconsistent beliefs) is a necessary condition for a full-blooded case of self-deception. On the other hand, the metaphorical approach is to examine self-deception as a form of

irrationality, and it is only the irrationality itself that matters, without the self-deceiver ever having to accept the falsehood of the favoured belief that he manufactured for his own satisfaction.

My approach will be to base self-deception on the model of other-deception, but on a sliding scale from "weak" deception (in which the opposite belief is not necessarily accepted, or in which the true belief is either avoided or, because of the flexibility of the ways in which the evidence can be interpreted, disguised) to "hard" deception (in which we do have to deal with two opposing beliefs or conclusive evidence). I have shown that there are "weak", as well as "hard", cases of other-deception (p.14) and these, are I think, in a general way analogous to "weak" and "hard" cases of self-deception. Perhaps it won't be necessary to adopt the "all-or-nothing" approach, to deny that literal self-deception is possible or to state that avoidance of the truth constitutes self-deception. In Chapter 4 I shall show how avoidance or disguising of the truth in acceptable terms is a <u>form</u> of self-deception, but in the later discussions on the various Theories of Mental Systems, which concentrate mainly on the possibility of "hard" cases, I shall show that if deception is seen as the holding of two contradictory beliefs, literal self-deception is possible. However, I want to show that it is misleading to claim that the genuine, full-blooded cases of self-deception are <u>only</u> those that parallel the epistemology of "hard" cases of other-deception. The meaning of "literal" self-deception more often than not refers only to the problematic and paradoxical case of aB(p+-p). However, I want to argue that full-blooded self-deception can also refer to aBp+aB-p. Martin, who views self-deception as a set of related phenomena which are analogous to other-deception in most but not all ways, uses the example of "self-taught" to illustrate the general, but not strict, analogy

between "self-deception" and "other-deception". He claims that the literal sense of "self-taught" is also paradoxical.

> "Teaching requires knowing something and being readily able to become explicitly conscious of it, whereas learning something entails just being ignorant of it. In order to be self-taught, a person would have to know and not know the same information and be readily able and not readily able to call it to consciousness. Even if this occured within a single person having a split personality, it would not, strictly speaking, be one self teaching itself. Hence, the idea of one person being teacher and student with respect to the same information is incoherent, and self-taught individuals could not exist." (Martin, 1986, p.20)

The flaw in the above argument arises through a too rigid analogy in a too-limited situation of "taught-by-others" being applied to "taught-by-self". The model of teaching others is helpful in limited respects to thinking about self-taught, because both activities are concerned with intentional action directed towards the acquisition of new knowledge. However, the helpful interpersonal model does not offer a complete understanding of the literal or standard meaning of self-teaching.

Even if we were to adhere strictly to the literal sense of self-deception in the limiting case and to interpret the phenomenon in terms of a split self, where one self deceives another, we are still not using the term literally, unless we restate it as "selves-deception". Furthermore, "when we speak of somebody's having more than one self, this is quite clearly metaphorical... When one such self deceives another, then, this is still not literal self-deception: while 'deceive' can be literal, 'self' becomes a metaphor." (Haight, 1980, p.36)

Pears, in Motivated Irrationality, suggests that if the investigation of

self-deception starts from the apparent implications of the concept, these will have to be checked against the actual use of self-deception, and if the investigation starts from the actual use, the possibility exists that some of the applications may have to be adjusted so that they are more in keeping with the real implications of the concept.

> "We cannot take the concept of self-deception and assume, with many psychologists, that it must be in good working order because there is general agreement about its application. But equally we cannot take the apparent implication of this concept, that a person deceives himself with full knowledge of what he is doing, and argue, on the side of many philosophers, that, since this is impossible, there is no such phenomenon as self-deception. What is needed is a balanced judgement and a due regard for the claims of denotation and connotation."
> (Pears, 1984, p.2-3)

Even if the claims of denotation and connotation are noted, self-deception still remains an irritating concept. Its denotation is far from clear, with vague borders between self-deception and, say, wishful thinking or self-deceptive akrasia. If its connotation is taken literally, the chances are that self-deception is not possible and it cannot, therefore, really have any denotation. Even if the word "self-deception" with its strict and problematic connotation did not exist in our vocabulary, we would still be faced with the problem its actual denotation raises. It seems as though the problem goes deeper than the paradoxes of self-deception to the various forms of the paradox of irrationality. For Pears, the deep issue of self-deception is the problem of irrationality: how can a person persuade himself that something is the case when all his evidence points to the opposite conclusion? The denotation of self-deception can include forming a belief based purely on desire, to forming a belief against the available evidence, to believing the conjunction of two logically incompatible propositions——

the limiting case of self-deception which is necessary if the connotation is to be interpreted literally.

Cases of other-deception were placed on a sliding scale in order of increasing severity, from withholding evidence to fabricating evidence. Self-deception displays the same characteristic of developing complexity. Pears distinguishes four grades of self-deception, from "weak" to "hard". (1984, Chapter V). The sliding scale distinguishes between avoiding or distorting or biassing already available evidence that may lead to a false but welcome belief (or for that matter, which may lead to the avoidance of a damaging but true belief) and manufacturing or fabricating a belief that goes <u>against</u> evidence already collected. Furthermore, I want to distinguish between evidence that allows a certain amount of "latitude" of interpretation and conclusive evidence (e.g. mathematical evidence) that rules out an appeal to various possibilities.

First of all, then, at the bottom of the scale is the "weak" case in which there is balanced evidence for <u>p</u> and not-<u>p</u>, and the self-deceiver accepts <u>p</u>. The person does appreciate the fact that he has evidence for both <u>p</u> and for not-<u>p</u>, but this does not mean that he is entertaining a contradiction. He may, of course, be experiencing a conflict if the evidence for the two different beliefs are close to being in balance. For example, Tim is an artist who has submitted his paintings to the Academy for acceptance and approval. If the Academy accepts his paintings, Tim will enjoy fame and public acclaim something he dearly desires. During the examination, one judge looks gloomily at the canvases, purses his lips and gives a decisive shake of the head. It is clear that he does not approve. From this evidence Tim forms the belief, "My painting is not accepted", (not-<u>p</u>).

However, the other judge smiles enthusiastically, rubs his hands together and nods encouragingly at Tim. From this evidence Tim forms the belief, "My painting is accepted." (p). There is equal evidence for both beliefs, but because Tim wants his painting to hang in the Art Gallery he firmly holds the latter belief, p, based on his desire for the acceptance of his art. This is still a case of irrationality, albeit a weak one, for a rational man would suspend belief. The self-deceiver, however, believes p firmly because he wants it to be the case that p. Tim forms a belief under the influence of a wish.

The above case merges with a case of wishful thinking (for the distinction between the two phenomena see Chapter 2), and in order to get a clearer case of self-deception, the irrationality will have to be increased, by supposing that his evidence points to the unwelcome belief. Here the person is faced with inductive evidence for not-p, but he accepts p. For example, when Tim meets with the members of the Art Academy there are three judges looking at his painting. He knows that in order for his work to be accepted, at least two of the judges must approve. However, judging from the disapproving looks on two or even all three of the judges' faces, Tim realizes that his chances of acceptance are virtually nil. However, Tim gambles on the slim chance that they might accept his painting, might change their minds when he has left, or might be putting on a public facade (they are, in fact, very enthusiastic about the painting but are skilfully hiding their emotions). The evidence is not totally conclusive and Tim, irrationally, stakes his chances on the latitude which the available evidence allows. (See Chapter 4 for further discussion of this form of self-deception.

The next step is to postulate a case in which the self-deceiver's irrationality is more extreme, because his premises are sufficient to

establish the falsehood of the favoured belief by deductive necessity. Here the self-deceiver is confronted with decisive evidence for not-p, but he accepts p. The conclusive evidence does not allow the "latitude" of interpretation in the previous case. The self-deceiver's adopted belief is logically incompatible with the evidence he possesses when he forms the belief. Here we are dealing with a more controversial case. The question raised is, "If someone's premises actually entailed the opposite conclusion to the one that he drew, how was he able to draw it?" (Pears, 1984, p.30) Since I am looking at intentional self-deception, the agent in this case is both competent and able to spot the irrationality. Pears cites the example of a competent accountant who, nevertheless, makes errors in his own favour when adding up his bank balance, but when he performs the same service for a friend, he makes no or few errors, or at least none that form a pattern with those he committed when he added up his own finances. (Pears, 1984, p.30). This sort of thing does happen, but by citing an example is in no way an indication that the problem of how it happens has begun to be solved. Pears suggests than an explanation would have to start from the fact that in a complicated case, it is far more difficult for the agent to achieve the same sort of commanding view of his premises and conclusion than he can achieve in the case of a simple contradiction. (Chapter 5 and 6 will deal with these issues).

And at the top of the scale is the most limiting (and some would say, impossible) case of self-deception: he knows that not-p, and yet the self-deceiver accepts p. He believes or even knows that something is not so, and yet he adopts the belief that that very thing is so. If the connotation of "self-deception" is taken literally, then this kind of case would involve the agent's believing two contradictory propositions and his believing the conjunction of two logically incompatible propositions. (Chapters 5 and 6

will look at whether this is at all possible). Since this limiting case at the top of the scale happens to be paradigmatic of the paradox of self-deception, it is this case which has captured the attention of philosophers.

* * * * * *

In this chapter I looked at the various definitions of "deceive" and how the different meanings give rise to different emphases. If "deceive" is taken as "cause to believe falsely that p", we can include cases of unintentional deception. However, the philosophically interesting case is that of intentional deception which rests on the stricter definition of "deceive", viz. "getting someone to believe something one knows to be false". And it is this form of deception, and thus self-deception, which will be investigated — the case which involves knowledge and intention (as necessary in the process of self-deception) and effect (as necessary for the state of successful self-deception).

The purpose of this chapter is to look at the central paradoxes embedded in the concept of self-deception when it is based on the model of other-deception, and the problems involved in an investigation of the notion of self-deception. The paradoxes are usually approached in two different ways: Firstly, there is the opinion which claims that self-deception can be treated only as being metaphorical to other-deception. By abandoning the model of other-deception, these theories of self-deception circumvent the paradoxes. Literal self-deception, therefore, is not possible and the phenomena which we falsely call self-deception are really cases of irrational belief-formation, rationalization, failure to focus, etc. Since these cases very often don't include inconsistent belief, or strong counter-evidence,

they are labelled as "weak" self-deception. These theories, however, fail to give an adequate account of how "hard" self-deception is possible, i.e. the case which involves inconsistent belief and conclusive counter-evidence. A second approach to the paradoxes is to confront them squarely by claiming that full-blown self-deception is possible. In order to explain how "hard" self-deception is possible, these theories appeal to the assumption of mental systems, i.e. the division of the psyche into different mental groups. I am, however, going to follow a third approach which, instead of an "either-or" choice, is going to incorporate both the above opinions. There are degrees of self-deception, ranging on a sliding scale from "weak" to "hard" cases. This does not mean that the model of other-deception is abandoned. Indeed, the analogy holds for there are also varying degrees of deception in inter-personal situations, from a withholding or manipulation of evidence to the fabrication of false evidence.

Using the model of other-deception leads to three main paradoxes. There is, first of all, the epistemological paradox which looks at the problem of holding inconsistent beliefs. I shall show later that not all cases of inconsistent belief are, however, cases of self-deception. What is needed is an intentional element, the agent intends to deceive himself. This leads to the second difficulty, i.e. the psychological paradox which looks at the problem of the self-deceiver's necessary awareness and unawareness of what he is doing. I shall have to confine my study to the solutions of the epistemological and psychological paradoxes, for the scope of my study does not allow enough room to include a discussion of the ethical paradox. I have merely noted this moral paradox in order to illustrate the complexities housed in the notion of self-deception.

Notes

1. E.g. Kipp (1980) and Paluch (1967)
2. E.g. Haight (1980) and Miri (1973)
3. E.g. Bach (1980) and Demos (1960)
4. E.g. Sartre (1958)
5. E.g. Fingarette (1969) and Martin (1979)
6. E.g. Demos (1960) and Butler as quoted by Gardiner (1970)
7. E.g. Kipp (1980) and Elster (1979)
8. There are, of course, theories which do not grant this assumption so readily. Bach (1980), for example, states that self-deception entails avoiding the unpleasant thought p, without necessarily forming the opposite belief not-p, and Mele (1983) regards the self-deceiver as manipulating data to favour his belief p, without our having to suppose that he once (or still) believed that not-p.
9. E.g. Bach (1980)
10. E.g. Mele (1983)
11. E.g. Kipp (1980)
12. E.g. Gardiner (1970)
13. E.g. Canfield and Gustafson (1962)
14. E.g. Penethum (1964)
15. E.g. Pears (1974)
16. E.g. Demos (1960)
17. E.g. Miri (1973)
18. E.g. Davidson (1982)
19. E.g. Fingarette (1969)
20. E.g. Elster (1979)
21. E.g. Radden (1984)
22. E.g. Foss (1980)
23. E.g. Hanson (1986)
24. E.g. Haight (1980)
25. Cf.Miri (1973) who adds the condition that B must recognize A's intention that B must believe that the statement is true. This apart from seeming superflous, does not, however, explain cases of deception in which A states the truth but in such a cynical or joking manner so that B will not accept it as the truth.
26. Brooks (1986, p.272) points out following Plato that (p+-p) is not necessarily an outright contradiction. We can have "John is both fat and not-fat". John may have a fat stomach, but his legs are thin, or John may be too fat for a jockey, but is thin for a member of his corpulent family. The term "fat" is, of course, flexible and it may be argued that a definite term like "dead" in "John is dead and John is not dead" will lead to a contradiction. However, I can reply that John is dead but his memory is kept alive by friends and family.

27. Priest (1986) argues that it may be rationally acceptable to believe in a contradiction. He cites Heraclitus, Plotinus, Nicholas of Cusa, Hegel and Engels as great rational thinkers who have consciously believed in explicit contradictions. "That a person may sometimes be able to accept a contradiction rationally, and that there is nothing in the domain of formal semantics ever to stop a person accepting a contradiction, I do not dispute." (p.111) What he does, however, reject is that a person can always accept a contradiction rationally. He concludes that when evidence and argument build up, it may no longer remain rationally possible to believe in a contradiction. And it is exactly this stage which is of interest in the study of self-deception: the rational impossibility of clinging to a belief that is faced with overwhelming counter-evidence.

28. See Davidson - (1970), (1982), (1985)

CHAPTER 2

RELATED IRRATIONAL PHENOMENA

Before going on to discuss what self-deception is, I want to discuss what self-deception is not. The following phenomena are like self-deception, but are not the same as self-deception. Indeed, they may shade into self-deception, or self-deception may grow out of them, but in each case there is a reason for distinguishing that case of irrationality from a case of self-deception proper. In this chapter I shall try to isolate the peculiar characteristics of self-deception which distinguish it from other forms of irrationality.

The first phenomena from which self-deception must be distinguished is that of intellectual incompetence. In the discussion of the psychological paradox in the previous chapter, I touched on the necessary condition of intellectual competence for a case of intentional self-deception. The paradox arises because the rationally competent agent intentionally and freely forms an irrational belief against the total weight of evidence available to him. The paradox would disappear if the person were hypnotized or drugged or failed to appreciate the impact of the counter-evidence.

If the agent did appreciate the impact of the counter-evidence, he would, as an intellectually competent person, be forced to form (to some extent, at least) a conflicting belief to the one he favours. If there is no conflict generated by opposing beliefs, there is no need for self-deception. In the general analysis of other-deception in Chapter 1, I noted that the first condition for successful intentional other-deception is that A makes

a statement to B which A knows or believes to be false. If A lacks the awareness(1) that what he says or believes is false, A is not intentionally deceiving, but is merely mistaken——a case of unintentional other-deception, e.g. when A inadvertently misreads a report in the newspaper. The intellectually incompetent agent who deceives himself (the meaning of "deceive" in this case refers to "causes himself to hold a false belief") fails to recognize the impact of the evidence or to see where the reasons point, whereas the intentional self-deceiver must know this only too well in order to know what to avoid or what to deceive himself about. The paradox embedded in intentional self-deception is that this very awareness and knowledge of the telling counter-evidence steers the agent on a path of self-deception——a path by means of which he can "conceal" or "disguise" this awareness and knowledge from himself. Moreover, not only does he have to conceal the knowledge of the import of the evidence from himself, he also needs to screen from himself the knowledge of the irrational causation of the favoured belief, as well as the knowledge that he is practising self-deception. The paradox of intentional self-deception, in other words, is that the self-deceiver deceives himself over the knowledge and awareness which is necessary for self-deception to take place. He devises a complex plan to embark on an irrational project, the project to hide or mark this awareness from himself. If he did not have the awareness, there would be no need for him to embark on a plan of self-deception, since there would be nothing for him to hide or disguise. Just how he is able to put this "rational" plan, aimed at irrationality, to work will be discussed in Chapter 5.

Without the agent's initial awareness of his own irrationality we are not faced with a paradox.(2) The intellectually incompetent person may fail to conform to <u>our</u> standards of rationality but his irrational belief may

seem perfectly "rational" to him. As I noted in Chapter 1, we can never have direct access to someone else's thought processes, but assume that they are similar to ours and we, therefore, impose general standards of rationality onto everyone. When someone's beliefs fail to adhere to these standards, we may label them as irrational. However, if the agent is not aware of the standards of rationality and not aware of the demands they make on his thinking, we cannot accuse him of being internally irrational, because he is not flouting any of his "rational" standards. In the case of the intellectually incompetent person it is this very intellectual incompetence which prevents him from becoming aware of his own irrationality.

However the case of the intellectually incompetent person is still troublesome. We can ask the question, "How can the person genuinely fail to notice the incoherence of his beliefs or genuinely fail to notice the impact of the counter-evidence?" But if the person never is aware (and never has been aware) that his beliefs are incoherent, he experiences no mental conflict, no incoherence or inconsistency in his belief-system.

> "The sort of irrationality that makes conceptual trouble is not the failure of someone else to believe or feel or do what we deem reasonable, but rather the failure, within a single person, of coherence or consistency in the pattern of beliefs, attitudes, emotions, intentions and actions". (Davidson, 1982, p.290)

Internal inconsistency is a necessary condition for irrationality within one person. Without internal inconsistency, the agent's action or belief may be irrational from <u>our</u> point of view, but it need not be irrational from the agent's point of view. (p.297). We would not, for example, speak of a three or four-year-old as being guilty of intentional

self-deception. The child does not yet fully know the kind of rational standards that are adopted in our ways of thinking about the world; he is not yet intellectually competent to apply those standards. In the child's thinking, contradictions are accepted without mental conflict, logic is something not yet grasped, and language is not yet sophisticated enough to form complex beliefs into propositions. The child's beliefs (for example, in the existence of fairies) and actions may seem irrational from the adult's point of view, but from the child's point of view there is no mental conflict between rationality and irrationality in holding those beliefs or performing certain actions. One must make "a strong distinction between lacking certain standards of reasoning and failing to apply them." (Davidson, 1985, p.141) A person who deceives himself into thinking that he is an excellent scientist will not be guilty of self-deception if his standards of excellence in science are low as a result of being simply ignorant of what good scientific work is or simply because he has had no competition and, therefore, no realistic evaluation of his own work. It is only when rational standards have been adopted by the person that we can speak of him as being self-deceived when he deliberately flouts those principles which he has acknowledged to himself. The first characteristic of self-deception then is that the agent must be intellectually competent and able to detect and experience mental conflict, brought about by the conflict between the demands of rationality and the desire for the irrational belief.

The second phenomenon I want to distinguish from self-deception is that of being in the grip of irresistable outside forces. I mentioned in Chapter 1 that the paradox of self-deception arises because the agent freely and intentionally forms an irrational belief. Furthermore, apart from being intellectually competent, he must also be able to avoid the

irrationality. "There is nothing surprising about a consciously irrational belief that is truly obsessional." (Pears, 1982b, p.162) If the self-deceiver has been overwhelmed by some alien passion, some powerful force which overcomes his reason, then it is implied that the self-deception is not intentional. Since the self-deceiver is not the agent of his action (including the action of forming the irrational belief), he cannot be held responsible. Davidson refers to this as the Medea Principle in which the rational faculty and better judgment of the person succumbs to the irresistable outside force.

> "If the agent is to blame, it is not for what he did, but because he did not resist with enough vigour. What the agent found himself doing had a reason——the passion or impulse that overcame his better judgement—— but the reason was not his. From the agent's point of view, what he did was the effect of a cause that came from outside, as if another person had moved him." (Davidson, 1982, p.295)

However, if self-deception is to take action in order to get myself to believe p which I believe to be false, then the question can be raised whether my going to a hypnotist or my taking an amnesia-inducing drug may then qualify as a sub-species of self-deception. A person could rely on an outside force or agency to instill in him a false, but welcome belief. For example, Tim, as a result of a very reticent personality, has no friends. His belief, "I am inhibited" is grounded on an objective evaluation of himself: his temerity to start conversations with strangers, his feeling of panic when someone approaches him, etc. He intentionally approaches a hypnotist to instill in him the opposite belief, "I am outgoing." Although Tim knows that at present there is strong counter-evidence to this belief, he hopes that this "implanted" false belief may influence his actions. He banks on the chance that through these extrovert and spontaneous actions he makes many friends who in turn give him the self-confidence he lacked before going to the hypnotist. He surmises that if

his actions do change as a result of the false belief, the future evidence (consequences of his changed actions) will allow him, quite rationally, to hold the belief that he is outgoing. Tim has fulfilled the conditions of intentionality (he deliberately approached the hypnotist) and he is also intellectually competent (he is all too aware of the irrationality of the implanting of the as-yet false belief). However, it is not clear whether Tim's subsequent action directed by the implanted belief is at all free or avoidable, the other necessary condition for self-deception proper. Although Tim intentionally approaches the hypnotist, it is the hypnotist who does the actual "implanting" of the false belief; it is he who manipulates the mechanisms of Tim's mind. Much has been written about the force of post-hypnotic suggestions, whereby it is clear that the subject, even after coming out of the hypnotic trance, feels compelled to perform a certain action which was suggested to him while he was hypnotized. These kinds of cases are problematic, but I would not call them paradoxical. In cases of self-deception proper the agent's rational faculty is defeated by that very rational faculty itself or by a wish. There are no intermediate "stage props" or outside forces which help with or facilitate the usurption of reason. Self-deception is not an ordinary case of irrational thinking. Irrational thinking may be caused by fatigue, shock, alcohol, brain damage or whatever, but self-deception is caused by the agent himself without the help of outside forces.

Similar to being in the grip of an irresistable outside force, is being deluded, that type of delusion that is typical of insanity. The schizophrenic who hallucinates and is convinced that he is seeing monsters, in the teeth of evidence, may be self-deluded in the sense that he has beliefs that are false and unreasonable, but we would not accuse the schizophrenic

of practising self-deception. The deluded schizophrenic experiences no conflict; there is no struggle between the demands of his rational standards and his delusions; there are no contradictory beliefs as there are in self-deception. He seems to be the helpless victim of some powerful forces, whereas the self-deceiver is seen to exercise some control over his self-deception; he intentionally and freely brings about his own delusion. Haight distinguishes self-deception from delusion by attaching an "air of responsibility or choice" to self-deception, which delusion lacks. (Haight, 1980, p.3) The self-deceiver "seems to choose not to know the truth". (p.4). In cases of intentional other-deception, the deceiver knows that he is deceiving and that he is responsible for the deception. This air of responsibility for the deception must, therefore, also be present in cases of intentional self-deception, since I am basing my notion of self-deception on that of other-deception. "If the false belief is brought about by brainwashing, post-hypnotic suggestion, or whatever, then we would not call the believer self-deceived, for responsibility for the false belief would not be present." (Foss, 1980, p.242) Moreover, self-deception is not merely clinging tenaciously to an unwarranted belief despite the recognition of counter-evidence. This is rather a case of sheer stubbornness. Self-deception, on the other hand, is different from pigheadedness in that the self-deceiver freely and intentionally chooses to employ a process of reasoning which "reconciles" the counter-evidence with the favoured belief. The agent himself is responsible for the employment of the process aimed at overthrowing rationality. So, a second characteristic of self-deception proper (as opposed to unintentional self-deception) is that the agent freely and intentionally forms his own irrational belief which gives rise to mental conflict. The agent is directly responsible for bringing about his own deception.

The third distinction I want to make is between self-deception and hypocrisy. Because both deception and hypocrisy are forms of pretence, it might be tempting to think that hypocrisy is a species of deception. This is, however, not necessarily so. Martin points out that a person can remain a hypocrite, without deceiving himself or anyone else. Martin notes (1986, p.44) that Senator Joseph McCarthy remained a self-righteous, pompous, intolerant hypocrite even after everyone saw through his hypocritical posturing. Hypocrisy is a form of pretence, usually a pretence of being better than one really is. There is no stipulation that this pretence necessarily leads to actual deception. The hypocrite may pretend to be better than he really is, without ever believing that he actually is better.

In hypocrisy, A may act as though he believes p, but he can never really believe p. "Hypocrites are self-aware in that they acknowledge to themselves their pretence, intentions, motives and strategies." (Martin, 1986, p.46) The difference between a hypocrite and a self-deceived person is that in the former case the person is only the agent in the pretence, whereas in the latter case the self-deceived person is both agent and victim of his own pretence; he is taken in by his own deception. The self-deceiver comes to believe what he knows to be false, but when a person pretends to himself it is not true that he believes that p; he only make-believes that p; he only behaves as if he believed. When a hypocrite starts believing that what he is pretending about is true, he is tending towards self-deception, or is guilty of what Butler calls "inner-hypocrisy". The hypocrite is engaged in dishonesty with himself while pretending to himself that he is not engaged in dishonesty. When the hypocrite reaches this stage, he is guilty of self-deception in that he is taken in by his own pretence. So another distinctive feature of

self-deception is that the self-deceiver does not merely behave as if he believes that p, he actually does believe that p.(3)

Another phenomenon that is closely related to self-deception, especially in the philosophical investigations of irrationality, is akrasia. Akrasia can be loosely translated as weakness of will, but to highlight the equally paradoxical core of akrasia, Pears refers to it as "action done knowingly against one's own better judgment." (Pears, 1984, p.15) The akrates holds certain beliefs and makes a value judgment, and the next moment he goes against his own judgment. He has reasons for judging a certain cause of action better than another, and then freely and intentionally goes against these, and follows another course of action. The self-deceiver has certain evidence and facts from which he forms a certain belief and then, freely and intentionally, he goes against the evidence and forms a conflicting belief.

I shall just look at some similarities between self-deception and akrasia, but will then show how they differ from each other. The similar paradoxical core of both concepts seems to be that the irrationality is caused by the same culprit, i.e. a rebellious wish or desire in the agent, which gives rise to either irrational action or irrational belief-formation, or both. Self-deception, as irrational belief-formation, shares distinctive features with irrational action or akrasia. Pears even goes as far as saying that the problem of irrational action contains the problem of irrational belief-formation within itself. (Pears, 1984, p.25) He offers two reasons for this. First of all, the formation of an irrational belief is often a kind of action and secondly, irrational actions are often made easier by biasing beliefs. Audi too stresses the reciprocal relationships between self-deception and akrasia. He notes that we "become self-deceived partly

or largely through actions, and act as a result of being self-deceived." (1982, p.148) So akrasia can contribute to producing self-deception, and self-deception can help to perpetuate akrasia (e.g. by evading telling evidence). What produces the action or biases the belief is usually a rebellious desire. Pears illustrates the point with an example of a guest at a party who desires a third double whisky, but concludes, rightly so, that it won't be all right for him to drive home after six measures of whisky. However, the desire for the third drink biases his reasoning. Thus by removing an intellectual obstacle to its own fulfilment, the desire leads to the guest's concluding, against the weight of his evidence, that a third drink will be all right. In other words, the rebellious desire causes the man to form the irrational belief that it is in order to drive home after three double whiskys. The desire, therefore, motivates the formation and acceptance of the irrational belief and this very same desire is then fulfilled by the action, i.e. the taking of the third drink. Both the biasing of beliefs (self-deception) and the irrational action (akrasia) are motivated by the same desire; both self-deceiver and akrates act intentionally on a favoured reason.

A further similarity between, in this case, "weak" self-deception and akrasia is that they both operate in situations with latitude, situations in which there is either balanced evidence for p and for not-p, or inductive evidence which tends towards not-p (Cases (1) and (2) on Pears' sliding scale in Chapter 1). Davidson (1985, p.139) generalizes self-deception as a common situation in which the weight of the available evidence seems to point to the truth of some proposition, and the agent is inclined to treat this proposition to be true more likely than not. In a situation with latitude, the agent does not have absolute certainty of the truth of a proposition, nor absolute certainty that its negation is

also true. His desire for the favoured belief will incline him to seek evidence or emphasize certain facts which will support the favoured, but suspected false, belief. However, to operate in a situation with latitude is not to rob the agent's belief-formation or action of irrationality. The latitude excludes absolute certainty about the truths of opposing propositions, but it still entails a context of conflict. If the latitude in interpreting certain ambiguous evidence or values did away with conflict, there would be no irrational belief-formation or irrational action. "The existence of conflict is a necessary condition of both forms of irrationality." (Davidson, 1985, p.140) "Weak" self-deception involves forming an irrational belief in the face of conflicting evidence, whereas akrasia involves an irrational action in the face of conflicting values. "Weak" self-deception is thus when the agent is faced with conflicting evidence for conflicting beliefs, having no certainty about the truth of either, but his desire for the favoured belief biases the evidence so as to seem to support the belief. Davidson compares this characterization of self-deception to that of akrasia or what he terms "weakness of the will":

> "Weakness of the will is a matter of acting intentionally (or forming an intention to act) on the basis of less than all the reasons one recognizes as relevant. A weak-willed action takes place in a context of conflict; the akratic agent has what he takes to be reasons both for and against a course of action. He judges, on the basis of all his reasons, that one course of action is best, yet opts for another; he has acted 'contrary to his own best judgment'." (Davidson, 1985, p.139)

So, even though self-deception and akrasia can appeal to latitude when interpreting ambiguous evidence or values, both are still irrational in that they operate in a context of conflict——between the demands of reason for the rational belief or the final value judgment, and the desire

for the irrational belief or action——and that in both cases irrationality triumphs.

From the above, it may seem that self-deception and akrasia, although irrational, are not so irrational as to give rise to inexplicable paradoxes. After all, the akrates who takes a third whisky against his own better judgment is not guilty of any logical inconsistency even though, according to Davidson, his unconstrained intentional action indicates that he judged it better to take another drink. "To explain his behaviour, we need only say that his desire to do what he held to be best, all things considered, was not as strong as his desire to do something else." (Davidson, 1982, p.297) Although it is irrational for the agent to go against his better judgment, he is not guilty of internal inconsistency. We cannot yet accuse the akrates who judges, "It would be better not to take another drink (first value-judgment); but another drink is desirable and I shall take a third whisky (second, perverse value-judgment)" of logical inconsistency.

According to Davidson, what is needed to reduce the akrates' irrationality to logical inconsistency is for the akrates to adhere to a second-order principle, a higher-order judgment that he ought to act on his first value-judgment ——"I ought to act on my own better judgment, what I judge best or obligatory all things considered." Therefore, for the guest to be guilty of internal inconsistency, the kind that gives rise to problematic paradoxes, he would have to accept the following two principles: "I judge it best not to do a, but a is desirable and so I shall do a", as well as "I ought to act on my own best judgment". If that judgment is made, that he ought to act on his best judgment (i.e. the judgment that it is best not to take the drink) and if he then goes ahead and does

take the third whisky, then the akrates' irrationality turns into "pure internal inconsistency".

> "A purely formal description of what is irrational in an aktratic act is, then, that the agent goes against his own second-order principle that he ought to act on what he holds to be best, everything considered. It is only when we can describe his action in just this way that there is a puzzle about explaining it. If the agent does not have the principle that he ought to act on what he holds to be best, everything considered, then though his action may be irrational from our point of view, it need not be irrational from his point of view." (Davidson, 1982, p.297)

Now, according to Davidson, just as the akrates is guilty of internal inconsistency only when he sins against his second-order principle, so the self-deceiver is guilty of internal inconsistency only when he sins against a similar second-order principle, the normative principle of the requirement of total evidence for inductive reasoning. (Davidson, 1985, p.140) When the agent is deciding among a set of mutually exclusive hypotheses, this requirement enjoins him to accept that hypothesis most highly supported by all available relevant evidence. To illustrate Davidson's point, I want to return to the example I used in Chapter 1 to illustrate a "weak" case of self-deception, the example of the artist, Tim, who submits his painting to the Academy for inspection. Although two of the three judges look decidedly grim and critical, Tim banks on the chance that they may change their minds, and he, therefore, forms the belief that his painting is accepted. It is irrational for Tim to do so, for his desire to have his painting accepted biases his reasoning and allows him to form a belief which he would not have formed in the absence of that desire. Although Tim is guilty of irrationality, we cannot, according to Davidson's stipulation of a second-order principle, accuse

him yet of internal inconsistency. For Tim to be guilty of that, he will have to hold the second-order principle of: "I ought to accept the hypothesis or belief most strongly supported by all the available evidence." Tim knows that all three judges must approve his painting before it is accepted; he is aware of the very strong evidence which leads him to form the belief that his painting is not accepted; yet his desire for fame leads him to form a belief that his painting is accepted, based on only part of the evidence (the sole enthusiastic judge); however, he also holds that he must accept the belief most strongly supported by the evidence. If all the above apply to Tim and if he yet freely accepts only the favoured belief that his painting is accepted, Tim is guilty of "pure internal inconsistency". The person "whose thinking does not satisfy the requirement of total evidence may be irrational by one person's standards but not (if he does not accept the requirement) by his own standards". (Davidson, 1985, p.141) So, like the akrates, the self-deceiver's belief may be irrational from our point of view, but it does not necessarily mean that the self-deceiver is internally inconsistent, i.e. irrational from his point of view. "All genuine inconsistencies are deviations from the person's own norms. This goes not only for patently logical inconsistencies but also for weakness of the will ... and for self-deception". (Davidson, 1985, p.142)

The similarities between akrasia and self-deception are, thus, that both agents act intentionally on a favoured reason; although latitude may be present, both agents operate in a context of mental conflict; for the irrationality of both self-deception and akrasia to harden into pure internal inconsistency, both agents must accept a second-order principle. It may seem that the one is merely a sub-species of the other, but there are important differences which distinguish them from each other. For

Pears, at least, the irrationality of an action that fails to conform to the agent's better judgment is very different from the irrationality of perverse belief-formation:

> "[The irrationality of perverse belief-formation] involves, in the limiting case, believing something impossible, but the irrationality of perverse action is quite different. The action is irrational because it is unreasonable and it is unreasonable only in the sense that it does not obey reason. It is not an element in the agent's reasoning and so it cannot be faulted in anything like the way in which we fault the conclusion of a piece of theoretical reasoning or an inconsistent set of beliefs. The agent's reasoning terminates with his singular value-judgment about his particular predicament and then his action is irrational only in the sense that it is not ruled by this edict of reason."
> (Pears, 1982b.p.167)

For Pears then, in self-deception there is a perversion within the reasoning process, whereas in akrasia the action does not conform to the final value-judgment reached by rational and logical procedures. Pugmire notes (1982. p.185) that Pears favours an account of akrasia in which the agent deviates from the choice dictated by his values and in terms of which the agent's deviation is seen as a rebellion against his practical reasoning, rather than an error within the agent's practical reasoning itself.(4) For Pears (1982b, p.176) it is not possible for a sane person consciously to form a belief that goes directly against a perceived fact. It is, however, possible for him consciously to perform an action that goes directly against his own final value-judgment. The agent may remember his value-judgment, be aware of the demand that it makes on his action, have no doubt about the action that would meet this demand, and yet perform the opposite action. The akrates, therefore, can in full awareness act directly against his final value-judgment, but it is most unlikely that the self-deceiver can in full awareness form and hold a belief that goes directly against the evidence. The akrates is, of course, acting

irrationally, but we cannot accuse him of internal inconsistency. It is only when he has accepted the second-order principle that the question of possibility of achievement is considerably curtailed.

Self-deception and akrasia often reinforce each other, but they are not the same thing. The most striking difference between the two is that the outcome of akrasia is an action, while the outcome of self-deception is a belief. "The problems raised by weakness of will differ distinctively from the problems raised by self-deception. The former problem is how a judgment can fail to lead to appropriate action. The latter problem concerns the way in which beliefs and judgments can be influenced by desires." (Watson, 1977, p.326) Another characteristic of self-deception then is that it is a perversion within the reasoning process which has as a result the formation of an irrational belief.

The last phenomenon I want to distinguish from self-deception is that of wishful thinking. Here, as opposed to akrasia, wishful thinking, like self-deception, has the outcome of irrational belief. In wishful thinking and in most cases of self-deception the belief is caused by a wish and it is this that makes these two phenomena irrational. A desire or wish is not a justified cause for a belief, except for beliefs such as the belief that you have that desire or wish. Szabados in his paper Wishful Thinking and Self-Deception examines the two related irrationalities of the self-deceiver and the wishful thinker in detail.

> "Both hold the belief they do hold largely because they want to believe that (p). The truth of the belief would make them happy. They both have a personal stake in what they believe. In the absence of the motives that we ascribe to them, neither would believe what they presently believe. Both have motives, subjective reasons, for holding the belief that they do hold." (Szabados, 1973, p.204)

Wishful thinking is a case of believing something because the wishful thinker wishes it to be true. The wish is a reason for believing that p and it provides the agent with a motive for acting in such a way as to promote having the belief that p. The wish, therefore, provides reasons of two kinds: firstly, the reason for believing that p and secondly, reasons for acting in such a way as to generate the belief that p.

> "The wish that p were the case ... can easily engender a desire to be a believer in p, and this desire can prompt thoughts and actions that emphasize or result in obtaining reasons of the second kind. (There is nothing) necessarily irrational in this sequence. An intentional action that aims to make one happy or to relieve distress is not in itself irrational. Nor does it become so if the means employed involve trying to arrange matters so that one comes to have a certain belief." (Davidson, 1985, p.143)

So, even though the belief is caused by a desire, the belief and the other consequences of that desire are not necessarily irrational. Merely wishing something to be true is not on its own an obstacle to sanely believing it. For example, one of the most prevalent cases of wishful thinking is surely the wish that "X loves me". Emma's desire, that is the wish that Tim returns her love, causes her to want to believe that it is the case that Tim loves her. The negation of the belief in Tim's love would distress her too much, so she clings to the wishful belief which makes her happy. In other words, her desire for Tim's love is a reason for her belief that he loves her. Moreover, she may manoeuvre circumstances in such a way that may seem to support her favoured belief——she ensures that she always sits next to him at functions; she wangles invitations to his parties; she engages him in conversation at all opportunities, etc. The fact that they are always together may, in turn, provide her with a reason for believing that he loves her. There is

nothing necessarily irrational in Emma's cunning arrangement of matters. Emma is guilty of wishful thinking, but she actively tries to orchestrate circumstances in such a way as to provide rational support for her wishful belief. What would make it irrational is if there is evidence stacked against the belief——if, for example, Tim has stated on many occasions that she should stop pestering him.

In wishful thinking the person wishes that p were true and he believes that p, although he lacks good grounds for believing it. (Just how the wish that p causes the belief that p via the Freudian shortcut will be discussed in greater detail in Chapter 5). The wish causes the belief where there is no evidence for the belief or, at the most minimal counter-evidence. Bach states (1981, p.351) that wishful thinking need not involve any reasoning or semblance of reasoning. The wishful thinker imagines some state of affairs, likes what he imagines, wishes that it were the case, and supposes that it will come about. He does not try to justify this supposition, perhaps he is content with the lack of evidence. This would be a case of passive wishful thinking, where the desire is the reason for holding the belief, but there is no attempt on the part of the wishful thinker to act in such a way as to provide himself with a reason which "justifies" his holding the belief that p. Self-deception may start with the kind of wishful thinking of above, but what must be added to this case of wishful thinking to make it self-deception are grounds against p; the minimal counter-evidence grows into strong counter-evidence of which the person is aware and of the need to deal with it. Whereas the wishful thinker believes on too little evidence, the self-deceiver believes against over-whelming evidence. The wishful thinker does not actively pervert the rational procedures whereby truth and falsehood are established. He may actively try to orchestrate optimal conditions as

lending support for the favoured belief, but he still operates within limits of the rational procedures, exploiting them for his own use. However, as Szabados points out: "a crucial point of dissimilarity between wishful thinking and self-deceit is that in self-deceit the evidence is against the belief held. Once this is pointed out to the person involved; if he then proceeds to resist, by ingenious tactics, the natural implications of the evidence, we feel that he is self-deceived". (1973, p.205) I shall return to the example of Emma to show how wishful thinking differs from self-deception on the grounds of strong counter-evidence. Emma wishes that Tim loves her. Her wish causes her to form the belief that Tim loves her. This belief is irrational in that it is caused by a wish and that there is no evidence on which the belief is based. However, there is no telling counter-evidence either: Tim is always friendly to everyone, including Emma; he does talk to her on occasion; he has never said that he does not love her. The absence of strong counter-evidence allows Emma the latitude in which she can form the favoured belief. It is, of course, irrational for Emma to misconstrue the absence of strong counter-evidence as an indication of supporting evidence —— the fact that he has never said that he does not love her, she concludes, must mean that he does love her. However, for Emma to be a candidate for self-deception, there is not merely an absence of strong counter-evidence, there must actually be strong counter-evidence of which she is aware: Tim has refused all her advances; he never returns her telephone-calls; he has even asked her to stop bothering him. The evidence is now too strong to allow her to hold, fairly comfortably, the belief that Tim loves her. The "ingenious tactics" to which Szabados refers are the various self-deceptive projects on which she may now embark. The rational person would, of course, at this stage abandon his belief, but the self-deceiver tenaciously clings to his belief, while deliberately

ignoring the strong counter-evidence or explaining it away. The self-deceived Emma may maintain that Tim does really love her, but he is too shy to show his real feelings; or she may blame Tim's antipathy on his parents who she claims have indoctrinated him against her because of her humble background; even when Tim announces his engagement to someone else, she may state that this is merely a ploy on his behalf to appease his parents, but that he will soon elope with her. The difference between the wishful thinker and the self-deceiver is that the wishful thinker, in the absence of evidence, has only one belief, the irrational but pleasing belief, "X loves me". The self-deceiver, however, is faced with strong evidence which forces her, as a rationally competent being (the case of irrationality I am interested in is, after all, that of the person who is competent and able to avoid the irrationality) to form the belief, "X does not love me". But she perversely holds on to the favoured belief, "X loves me". Whereas in wishful thinking the agent has one belief, the self-deceiver is faced with two contrary beliefs. Wishful thinking does not operate in a context of mental conflict whereas self-deception does; the conflict between the two contrary beliefs or the conflict between the favoured belief and the counter-evidence.(5) In wishful thinking there is no contradiction, whereas in self-deception there is a natural tension: "the sense of the evidence against p pulls one away from the deception and threatens to break down one's defences; the motivation and weakness that sustain the deception pull against one's grip on the evidence and threaten to overthrow one's perception of the truth". (Audi, 1982, p.140)

There are many cases of self-deception that have started as cases of wishful thinking which then grew "stubborn and perverse".(Haight, 1980, p.2) Wishful thinking shades into self-deception, with no clear line between

them. It can be hard to say at what point the facts begin to show not just a lack of evidence for p, or minimal counter-evidence for p, but constitute evidence against it. Or, as Davidson would put it, it would be difficult to tell just when the evidence against the belief is heavier than that in favour of the belief. Pears points out (1974, p.103) that in wishful thinking the wish does not flout the precept "Accommodate your beliefs to your evidence". What it does flout is a different precept of reason: "Get all the available evidence you need". As a rational being it may be a form of irrationality to ignore one's suspicion that there might be counter-evidence, but the person who does not collect all the available evidence is not being inconsistent. This view of wishful thinking overlaps with a large area of self-deception, namely those cases of "weak" self-deception in which the ambiguous evidence allows the person to ignore those facts which he suspects might tell against the favoured belief. He accommodates his belief to only part of the evidence——this is made possible by the latitude of interpretation which the evidence allows——and deliberately ignores or refuses to investigate facts he suspects to be counter-evidence. He is, of course, being irrational, but not yet internally inconsistent. It is only when he collects all the available evidence, and is aware that he is forced, by the overwhelming facts, to form a contrary belief and yet stubbornly clings to the favoured belief, that the irrationality deepens. This would be Pears' point of distinction between the wishful thinker and the "hard" self-deceiver. The self-deceiver no longer is able to accommodate his favoured belief to his evidence. A further distinction between the two is that at this point the self-deceiver is forced to hold two contrary beliefs, whereas the wishful thinker holds only one belief. And yet, for the self-deceiver to be guilty of internal inconsistency, he would, in addition, have to flout his accepted second-order principle that he ought

to accept that belief most strongly favoured by the evidence.

Davidson (1985, p.143) notes a further similarity between wishful thinking and self-deception in that both can at times be benign. The parent may think his child more intelligent than what the teacher or I.Q. test evaluate him as being. There are, of course, psychological results which support the finding that the child may in fact do better through being thought more intelligent. However, a case of self-deception will develop when the discrepancy between results and evaluation is just too great to sustain the belief in the child's genius. Davidson offers an example of "charitable self-deception aided by wishful thinking" in which a wife, in order to keep the family peace, may ignore the lipstick mark on her husband's collar. Both these cases of self-deception aided by wishful thinking have a motivational element at work: the parent hopes that his child is more intelligent than others think him to be and so motivates him to live up to this standard; the wife wishes to preserve the family stability. I mentioned earlier on that self-deception can at times be benign, but at times it can have also a destructive motivational element. It is this which is another difference between wishful thinking and self-deception. "In wishful thinking belief takes the direction of positive affect, never of negative; the caused belief is always welcome. This is not the case with self-deception. The thought bred by self-deception may be painful." (Davidson, 1985, p.144) Even if the depressed wishful thinker wants everyone to hate him and holds the erroneous belief that everyone hates him, he will still, however perverse it may seem to us, welcome the realization of this belief——it will make him happy if everyone were to hate him. The self-deceiver's belief, on the other hand, may be painful to that person. The wife who has deceived herself that her husband is being unfaithful may find "evidence" everywhere which confirms her worst suspicions.

Another characteristic of self-deception then is that the self-deceiver is faced with conflicting beliefs, but accepts the favoured belief even though there is strong counter-evidence. According to Pears, one of the most distinctive differences between wishful thinking and self-deception is the difference of the motivational aspect of all cases of wishful thinking and those cases of "cold" self-deception which are unmotivated by desires. All cases of wishful thinking are motivated by desires, whereas there are cases of self-deception in which the irrationality of the belief is due to an intellectual perversion of reason, unmotivated by desires. Although Pears makes much of these "cold" cases, I do think that these cases constitute only a small minority. However, in the next chapter I shall look briefly at "cold" cases of self-deception.

* * * * * *

In this chapter I have tried to draw certain boundaries around the concept of self-deception. It has proved to be a difficult task for I have found that self-deception, like a corpulent lady, cannot be forced into an all-confining corset, but bulges beyond the confines of the various theoretically imposed restrictions. On the one hand, self-deception shades into delusion in which there is a certain avoidance or perversion of reality. On the other hand, self-deceivers seem to be guilty of cunning pretence and hypocrisy. Self-deception differs from intellectual incompetence and yet self-deception itself involves the usurpation of rationality. Furthermore, self-deception is often used to facilitate akrasia and, conversely, akrasia is often called into the employ of self-deception. It seems as though wishful thinking is the start of many forms of self-deception, but it is difficult to identify the point at which self-deception ceases to be wishful thinking. Is it when there is strong counter-evidence, or is it

when the agent deliberately refuses to collect all available evidence? I do not pretend to have found all the answers to the many shifting subtle difficulties in defining a definite demarcation, separating self-deception from other related phenomena. In fact, I suspect that if a too definite border strictly separating self-deception from these other phenomena is established, there may be the danger of the theory not being in keeping with the actual phenomenon. Self-deception is, in its very execution dynamic, constantly shifting and adapting.

Notes

1. I have used the term "awareness" rather than "knowledge". "Awareness" allows more flexibility and latitude. It can incorporate all the notions of mere suspicion, belief and knowledge. Although "knowledge" is not the same as "certainty", it does, however, have a more rigid connotation than "awareness".

2. Non-awareness of the inconsistency of his beliefs or of the weight of the counter-evidence may, of course, be the result of successful self-deception. However, I hold with Pears that the self-deceiver's reason for initiating the project of self-deception is his uncomfortable suspicion, belief or knowledge of the inconsistency in his belief-system.

3. In Chapter 4 I shall discuss a case of self-deception which does not involve the self-deceiver's holding the belief that p; all he does is to avoid the sustained or recurrent thought that not-p. Although I do discuss it as a case of self-deception, I regard it as a peripheral case between self-deception and wishful thinking. The characteristics I am highlighting in this chapter refer to the typical cases of self-deception.

4. This differes from the views of both Aristotle and Davidson who see akrasia not as an action which fails to conform to the agent's final value-judgment, but who see the akrates as being guilty of irrational reasoning which leads to an irrational final value-judgment. The action conforms to the final value-judgment, but because this is in itself irrational, the corresponding action is also irrational. For Davidson, the akratic "fault" occurs within the agent's reasoning (it is incomplete) and not "between" his reasoning and his actions. This view brings the irrationality of akrasia much more in line with the irrationality of self-deception, since both are failures within reason itself. However, I wish to confine my study to self-deception proper, and wish to draw attention to similarities and differences with other related phenomena which demonstrate degrees of self-deception, without being enticed into a full discussion of those phenomena themselves.

5. It may seem that the conflict between the favoured belief and its counter-evidence necessarily entails a conflict between two contrary beliefs, the favoured belief and the rational belief based on the available evidence. However, as pointed out in note (8) of Chapter 1, there are those who hold that self-deception does not necessarily mean that the opposite belief is formed. This form of self-deception will be more fully discussed under the concept of "evasion" in Chapter 4.

CHAPTER 3

CAUSES AND STRATEGIES OF SELF-DECEPTION

In this chapter I shall first look at the causes of irrationality in general, and then more specifically at how these causes operate in cases of self-deception. In the previous chapter I noted some causes of irrationality, such as intellectual incompetence and being in the grip of some external force. Other causes of irrationality may include negligence, impetuosity, fatigue, etc., bringing about a case of inconsistency and irrational belief-formation. However, "it is not clear that there is a genuine case of irrationality unless an inconsistency in the thought of the agent can be identified, something that is inconsistent by the standards of the agent himself." (Davidson, 1985, p.138)(my emphasis). I have already mentioned in the previous chapter the distinction made by Davidson between external irrationality, an individual's failure to adhere to general standards of rationality, and internal irrationality, a failure of coherence and consistency within the individual's belief system. Cases of external irrationality appear to be irrational to us, and cases of internal irrationality are irrational by the agent's own standards. Both cases are obviously irrational, a failure of reason, but only the cases of internal irrationality are paradoxical, exhibit an inner inconsistency, such as in some cases of intentional self-deception. What is important to note is that not all cases of self-deception necessarily entail inner inconsistency. In Chapter 1 I mentioned cases of "weak" self-deception in which the agent relies on the latitude the evidence allows in interpretation to support his favoured belief. He is, of course, being irrational for the rational person would first garner more conclusive evidence before forming his belief. If the self-deceiver does not hold a

higher, second-order principle such as the requirement of total evidence, he does not violate his own standards and is, therefore, not guilty of inner irrationality. It is still a problematic case of irrationality, but the paradox has been considerably weakened. However, the case of self-deception that is of particular philosophic interest is, of course, that of internal irrationality. Furthermore, it must be noted that "weak" self-deception may also display inner irrationality. In "weak" self-deception the agent, for example, evades the unwelcome belief that not-p; the action is not in itself necessarily irrational, but if he holds the second-order principle, the requirement of total evidence for inductive reasoning, and still avoids collecting suspected damaging evidence, he is guilty of inner irrationality. So too the self-deceiver who has better reasons for believing that not-p, but accepts the belief that p, will only be guilty of inner irrationality if he, at the same time, holds the second-order principle that he ought to accept that proposition for which there are better reasons. This distinction between external and internal irrationality is of great importance, since some causes of external irrationality (such as incompetence and fatigue) are not causes of internal irrationality or necessarily of intentional self-deception. I shall examine other causes of external irrationality as listed by Pears (1984,ch.2) and will look at how these following causes can be operational in cases of self-deception.

The first cause is that of misperception. It is easy to see how someone's misperception of, say, a price-tag may lead to the formation of a false belief that the Porsche costs only R1,000. This is obviously a case of mistaken belief, but not of inner inconsistency. For there to be inner inconsistency there must be two (or more) inconsistent beliefs and a second-order principle which give rise to mental conflict. However, the correct

price-tag, i.e. R100,000 may never have been registered in the man's mind in the first place, and this would, therefore, be a case of mistaken belief and not of internal irrationality, since the correct price did not constitute "material already in the mind". (Pears, 1984, p.6)
This is a clear case of misperception of evidence caused by short sight. This type of misperception due to a physiological defect cannot be a cause of intentional self-deception, since in the case of a mistaken belief due to misperception there is no further input in the mind, which will challenge the person's erroneous belief. Since there is only the mistaken belief in the person's mind, there can be no mental conflict generated by an opposing belief. And in the previous chapter I have shown that mental conflict is a characteristic of intentional self-deception. If, however, the keen Porsche fan does not suffer from any physiological defect, and still misreads the price-tag because he wants it to be the case that the Porsche costs only R1,000 we have a case of motivated misperception, a case closer to self-deception and to which I shall return presently.

The next cause is that of forgetting which seems to steer us in the direction of problematic irrationality because the information has been registered in the mind at some stage. However, before this forgotten information can be used, either rationally or irrationally, it needs to be retrieved. If it is not retrieved, the information is not in the focus of attention and is, therefore, not available for processing. Penelhum notes that when the information is not registered in a person's mind or, in other words, if the person does not "know the evidence, we have not self-deception but ignorance". (1964, p.88) A case of genuine forgetting, then, is not an example of self-deception. The point I am labouring here is that genuine forgetting and misperception due to a physiological defect can be causes of mistaken belief-formation. However, the agent who forms

and holds these mistaken beliefs may be judged by others as being irrational, but he is not irrational according to his own standards. In other words, instances of incompetence such as misperception and forgetting may be causes of external irrationality, but are not causes of internal irrationality. There is no evidence registered in the mind to challenge existing beliefs, no evidence registered in the mind that leads to the formation of a contrasting belief, and, therefore, no mental conflict. That incompetence gives rise to irrationality is clear, but unmotivated incompetence (such as genuine forgetting or misperception due to a physiological defect) is not applicable to cases of intentional self-deception in which one of the characteristics of the self-deceiver is that he is a rational being who is both able and competent to detect and avoid irrationality.

However, what needs to be distinguished are cases of misperception of evidence caused by short sight or genuine bad memory and cases of deliberate misperception of evidence or intentional forgetting. The interesting question which now arises is what generates the failure to know or have the evidence. If the person deliberately ignores or is motivated to disregard or repress unwelcome beliefs, we are much closer to a case of problematic irrationality and self-deception. What seems to generate the irrationality in self-deception is a rebellious wish, which leads to a desire-influenced manipulation of evidence. Freud has shown that a wish can often manipulate reason, and by forwarding this view, he has overthrown the traditional view of reason as a completely independent force, stronger in some people than in others; as a force that allows no interfering with its inner working. Cases of motivated irrational belief-formation and action, such as deliberate misperception and forgetting, which are caused by a wish are termed "hot" cases of irrationality by Pears,

following a usage in psychology. In "hot" cases the agent's wish for something causes him to form an irrational belief through deliberate misperception or wilful ignoring of counter-evidence.

As I have discussed in the section on wishful thinking in Chapter 2, a belief that is caused solely by a wish is an irrational belief, for a desire is a completely inappropriate cause for a belief (except for beliefs such as the belief that I have that particular desire). In a "hot" case of self-deception, therefore, the cause of an irrational belief is the wish to believe that p. By forming an irrational belief, the self-deceiver typically relieves some of the stress of the painful thoughts caused by things beyond his control. Although motivated irrational belief-formation, or self-deception, poses no great problem to the practitioner, it does pose a problem for philosophy. Philosophically, it needs to be explained why the self-deceiver's immediate goal is a belief; a belief, moreover, that seems very likely to be false in relation to the available evidence. Surely the self-deceiver's ultimate goal is the real thing? The dying cancer patient does not have as goal the formation of the belief, "I am not dying of cancer", but wishes rather that it were the case that he is not dying of cancer. Pears asks the question (1984, p.11) of how anyone can aim at belief in the real thing instead of the real thing itself. The agent faced with unwelcome evidence has two choices: firstly, he can try to change the real world to such an extent that the altered world will produce the satisfying belief——this, however, normally proves too difficult a task, for how will the cancer patient make it the case that he is not dying of cancer?——or, secondly, if the task of changing the world is too difficult to accomplish, the agent's wish goes for the belief instead, that belief that will give pleasure, i.e. the belief, "I am not dying of cancer". Therefore, if to change the world to suit his favoured

belief proves too difficult, the agent can follow a short-cut and form the favoured belief without, or even against, the supporting evidence. The reason for forming the belief is his desire for the pleasure that the favoured belief will give him. When self-deception, like wishful thinking, is governed by the pleasure principle, it goes straight for the belief if the real thing is too difficult to achieve. In this case, when self-deception is caused by a wish, the self-deceiver then is a person "who wrongly believes something to be true which he would not have believed to be true in the absence of the particular interest in the matter concerned that he has". (Gardiner, 1970, p.242) Although this definition illustrates the motivation of the self-deceiver, it does not offer a clear-cut distinction between self-deception and mere wishful thinking. As I've discussed in Chapter 2, the wishful thinker too forms a belief because he wants the pleasure that holding the belief will give him, but only the self-deceiver's wish goes further: it motivates the person to ignore wilfully or to avoid deliberately strong counter-evidence. Because the world is too difficult to change, the agent's wish goes for the promotion and acceptance of the favoured belief instead. Pears notes that the wish may have different goals: its goal is either to form the favoured belief in order that it will give the agent pleasure, or it may have as goal simply to eliminate an uncomfortable belief, or its goal may be to make it easier for the agent to give in to a temptation (a desired <u>action</u>) by eliminating a belief that stands in the way. I shall examine the first two goals only, for the third leads into a discussion of <u>akrasia</u> which, although it shares many similarities with self-deception, is yet distinct from it as I have shown earlier on. I want to confine my study to self-deception proper. Not only are there different goals that the wish can have, but it can also operate either secretly or openly. It operates openly when the person acknowledges the preference for another goal, or it operates

surreptitiously by screening itself from the agent's consciousness. But the question of just how the wish operates will have to be suspended for now. I shall return to the actual mechanisms whereby a wish causes a belief in Chapter 5. What I want to point out here is the fact that a rebellious wish is a cause of self-deception.

In order to illustrate the above cause of "hot" cases of self-deception, I shall use Pears' example of a girl who has a lot of evidence that her lover is unfaithful, but she does not believe it. (1984, p.44) This constitutes a case of what people will term typical self-deception. Let me first show how misperception and forgetting in this case need not necessarily make the girl guilty of practising self-deception, even though she is holding an irrational belief in her lover's faithfulness. When she entered the restaurant where her lover was entertaining another woman, she may not have recognized him for she may have left her spectacles at home. Even though she still firmly holds the belief that he is faithful, she cannot be accused of self-deception, for the counter-evidence is never registered in her mind. On the other hand, she may have recognized him with the woman, but because her work schedule since the encounter has been so overloaded, she has genuinely forgotten that she had actually seen him with her rival. Again, she cannot be accused of self-deception, for the memory has not been suppressed deliberately because of its fearful implications, but it has truly slipped her mind. The information, although once briefly registered, is now forgotten. We, as spectators, perhaps assess her belief as irrational but we cannot assess it as part of an intentional project of self-deception: the girl is not experiencing a mental conflict between two contradictory beliefs, nor does she experience a mental conflict that needs to be resolved somehow between the evidence and her belief. There is no deliberate perversion within the reasoning

process since she has not registered the evidence as evidence of her lover's unfaithfulness.

But the more interesting case is, of course, if she does not want to acknowledge her lover's unfaithfulness, or if she wants to believe that he is faithful because it is too painful to face the opposite possibility. This is a "hot" case of self-deception. In the first case the wish has as goal the elimination of the uncomfortable belief or suspicion that he is not faithful. The wish can lead her to deliberately avoid evidence that may support this distressing suspicion, and by motivated evasion of damaging evidence the distressing belief now stands unsupported and is, therefore, eliminated. Another way in which the wish to eliminate the unwelcome belief operates is to suppress it into the unconscious, a form of motivated forgetting. On the other hand, the wish can have as its goal the actual formation of an irrational counter-belief, one that will give her pleasure. The pleasing belief that her lover is faithful is caused by the wish which perverts her reason and allows her to form the irrational belief.

In the above discussion, the motivation for self-deception is provided by a wish for some desirable goal. But there need not always be a desirable goal. Pears notes that self-deception can be caused also by fear or jealousy. This supports Davidson's observation, noted in the previous chapter, that self-deception need not always be benign, which distinguishes it from wishful thinking. These emotions often lead the self-deceiver to form unpleasant beliefs against the available evidence and the promptings of reason. It seems strange that the wish should aim at an intrinsically unpleasant belief, one that causes distress. Fear may make someone run away; the fear causes him to want to run away; he wishes to avoid an

unpleasant situation. Similarly, when jealousy makes someone retaliate, it causes him to want to do so; he wishes to eliminate a rival. It is easy to see how fear or jealousy causes people to want to act in a certain way, but it is difficult to see how fear or jealousy causes people "to want, in the ordinary open way, to form exaggerated beliefs," (Pears, 1984, p.43), moreover, an exaggerated belief that will bring about pain. Fear and jealousy can cause self-deception, but if self-deception involves a wish causing a belief, then it is difficult to see the justification for postulating a wish in the cases of fear or jealousy. Pears notes (1984, p.43) that "(i)f it does involve a wish in these emotional cases, it is not a wish that is felt by the subject. We would have to postulate that it is kept in the background and operates surreptitiously." The self-deceiver is not aware that the object of the wish is the formation of the intrinsically unpleasant belief. He is not aware of the operation of the wish but may, of course, be aware of the ulterior goal of the wish, i.e. his safety or the elimination of a rival. The first step that needs to be taken by the wish in order to achieve its ulterior goal is the formation of the necessary belief. But in cases of fear and jealousy-motivated self-deception, this does not happen. In these cases there is a wish for the ulterior goal "but nature takes over at this point and sets up an emotional programme that ensures its achievement. The plan is nature's and not the person's, and that is why the formation of the intrinsically unpleasant belief is not felt to be the object of the wish". (p.44). However, to shift the burden of the paradox from person to "nature" does not simplify the problem. At a later stage I shall return to this question of the self-deceiver's reliance on the "discreet operation" of the wish. What is of importance at this stage is that in cases of desire-motivated self-deception the wish for p has as its intermediate goal the belief that p, the "messenger with good news", when p is

too difficult to bring about in the real world. However, in cases of fear or jealousy-motivated self-deception, the person is not aware of the operation of the wish that forms the intrinsically unpleasant belief that p, but is aware only of the ulterior goal, a goal that will produce satisfaction.

So far it seems as though failures of rationality are produced either by a rebellious wish or by incompetence. Pears notes (1984, p.9) that another possibility has come to the fore in the last twenty years. Cognitive psychologists have devised experiments which show that, even when no wish is operating, a failure of rationality need not necessarily be produced by incompetence(1). Even though the agent is perfectly capable of processing the information correctly and is also aware of the principles for correct processing, he can still make errors in rationality. The cause for these errors in rationality is that "reason itself has certain bad habits that produce them". (p.9). It would, therefore, be a mistake to attribute a failure of rationality to either incompetence or a wish. Pears terms these errors which have a purely intellectual source as "cold" cases of irrationality. In desire-motivated or "hot" cases of irrational belief-formation the cause of an irrational belief is the wish to believe that p. A desire is an inappropriate cause for a belief (except for beliefs such as the belief that he has that desire) and it is, therefore, irrational. However, in "cold" cases of irrational belief-formation the cause of the irrational belief is generally appropriate, but the operation is faulty. The reasoning itself is pure, but its operation is incomplete. Thus, in "cold" cases, reason has it own perversions.

> "Just as Freud had shown that many faults attributed to incompetence or chance are really motivated, so too these experiments

> have identified a further range of faults that neither belong to the province of chance nor are the result of ordinary incompetence. For people make them without the incitement of any wish in areas in which they are quite capable of proceeding correctly and even understand the principles of correct procedure. Of course, we may, if we like, classify them as a special kind of incompetence, but the important point is that they are not the kind of incompetence that we attribute to a person who finds a task beyond him". (Pears, 1984, p.45)

An example of a "cold" case of irrationality would be a case in which the agent gives salient evidence more weight than it is worth, or he may not know how to deal with statistical evidence, or he may attribute a person's behaviour to a particular disposition when its real source is something quite different, or he may obstinately hold on to a hypothesis even though evidence is telling heavily against it. It may be argued, of course, that in this last case the person's wish that his hypothesis will be proved correct will motivate him to "misperceive" the evidence and will, thus, allow him to cling to the initial hypothesis. The scientist may wish to prove his hypothesis correct so that he will be respected for his intellectual astuteness. His desire for academic acclaim is a personal wish probably accompanied by emotion and, therefore, "hot". Perversions of reasons do not, therefore, necessarily exclude any wish. Once again, there seems to be no definite boundary and exclusions: rather than a clear-cut difference between "hot" and "cold" cases, there seems to be a gradation here.

> "There must be many cases in which they ("hot" and "cold" causes) co-operate in the production of error and some in which the co-operation is unnecessary, because each would have been sufficient by itself, in much the same way that a man facing a firing-squad can be killed by two simultaneous bullets in the heart". (p.8)

The interpretation of the desire-motivated belief of the scientist holding onto his hypothesis despite strong counter-evidence is applicable since there can also be a possible interpretation of the irrationality in terms of a "cold" cause. For example, a medieval astronomer forms the hypothesis that the earth revolves around the sun, even though the available evidence (Ptolemy's "established" epicycles, the daily observation of the "moving" sun, etc.) tells heavily against it. When one proof fails, he turns to another way of proving his initial hypothesis correct. When this fails he tries out yet a different experiment, but he has no personal stake in the outcome. But would we label this scientist as irrational? Is he guilty of a perversion of reason? Surely not.

This brings me to an interesting issue. Some philosophers(2), will argue that the medieval scientist may have been unreasonable, a man of blind faith, but certainly not deceived. "(T)he person who believes in the face of adverse evidence might just be right. If that is so, then he is neither deceived nor self-deceived." (Foss, 1980, p.238) However, self-deception and the self-deceptive project cannot be in a state of suspension until the success or failure of the belief is established. "Deception" describes also what someone is engaged in doing, whatever the outcome. The verb "deceive" can, thus, be used in two different ways: the one way will include the eventual success of what the deceiver is doing, but the other way has a more restricted application in which the implications of success are in abeyance. This second way focuses exclusively on the process in the deceiver's mind. So, whatever the outcome, the deceiver is engaged in a deceptive project. Here the success of the project is in suspension, for the deceiver may fail to put the belief across, or the belief he puts across may turn out to be true (unbeknownst to the deceiver at the time of his deception.) But he is,

nevertheless, intent on deceiving the other person. In this restricted sense of "deceive" are there then grounds for accusing the medieval scientist of being guilty of "cold" self-deception, even though his belief in the truth of the first hypothesis turns out later to have been justified? I think not because there is a shift in emphasis in the intention of the agent. The astronomer's intention is not to deceive others or to convince them to believe in the truth of his hypothesis. I think his intention is rather to pursue a cultivated hunch and to establish the truth of his hypothesis first for himself before trying to convince others. So it seems that the intention of doggedly trying to prove this hypothesis is the distinguishing factor between self-deception and blind faith (or scientific stubbornness). If the intention is to find proof to establish the admittedly then-implausible hypothesis, he does not seem to be guilty of self-deception. However, if his intention is to influence the thinking of his fellow scientists (or to influence his own thinking in a certain way as opposed to the objective pursuit of truth) regardless of whether the hypothesis turns out to be true or not, we may have a case of self-deception.

I want to return to the other "cold" perversions of reason: the agent's susceptibility to salience. The competent agent is led astray in that the evidence against the plausible belief is more vivid or salient than the evidence for it (or evidence for the implausible belief is more vivid than the evidence against it). Just as reluctance to abandon a first hypothesis does not always involve a personal wish, so susceptibility to salience as a "cold" perversion of reason need not necessarily involve an emotion. "It would be an obvious mistake to suppose that people attach too much weight to salient evidence because they prefer it, or prefer to be swayed by it." (Pears, 1984, p.10)

A third bad tendency that reason exhibits in the construction of beliefs is the habit of attributing a person's behaviour to a different disposition than the actual one. Pears (1984, p.46) notes that this intellectual perversion may take two forms. The first one is that of a mistaken attribution of a particular behaviour to a particular disposition. For example, Tim wants to buy cigarettes. He knows that the corner café stocks his brand and so he walks to the café and buys a packet. His reasoning and actions are rational. As a spectator, I may, however, erroneously attribute this behaviour to the above, rational disposition. Perhaps Tim is so distraught that he picks up a packet of cigarettes merely at random and is not even aware of the brand he has chosen. The real source of Tim's behaviour is something different to what I attribute his behaviour. I attribute rationality where no rationality exists and thus form the mistaken belief that Tim buys that particular brand because he likes it.

The second form of the error is the failure of the spectator to take account of the circumstances of the action. I may attribute Tim's action of buying that particular brand to an obvious disposition——Tim's preference for that particular brand, when in fact his action may have issued from a less obvious disposition. Perhaps Tim has secretly given up smoking, but he buys those cigarettes because he knows that I smoke that particular brand. Tim's action is quite rational, but I made the mistake of attributing it to a different disposition and, therefore, form the mistaken belief that Tim buys the cigarettes because he wants to smoke them.

A fourth form of a perversion of reason is that of impartial rationalization. The rationalization may be an honest attempt, free from guilt, shame or personal gain, to explain a certain act. This "cold" rationalization is

not powered by any personal motive, for the self-deceiver may rationalize his own behaviour with the same complete impartiality with which he rationalizes other people's behaviour, but which will lead him to form a mistaken belief. For example, the athlete whose prowess is failing may blame his poor performance on too many carousing late nights. When a fellow competitor also delivers an uncharacteristic poor performance, he may ascribe it to the same reasons as those for his own decline in achievement. In other words, he rationalizes his opponent's defeat in exactly the same way as his own——or as he would his own, to exclude the obvious possibility that he does so in order to give more credence to the rationalization of his own defeat.

It is clear how these "cold" perversions of reason may lead to the formation of erroneous beliefs and, also of irrational beliefs, but how do these various "cold" causes apply to self-deception? I shall return to the example of the girl who has a lot of evidence that her lover is unfaithful, but she does not believe it. When she recognizes her lover with another woman, she may mistakenly attribute his intimate lunch-date with a woman to a mere business committment, something which his job often demands. He is, in fact, exploiting her tendency to make this attribution error, and he is always careful to plan his illicit encounters over lunch-time so that, if confronted, he can appeal to the purely business requirements of his job. So, even though she is aware of his numerous lunch dates with the woman, she attributes his behaviour to the demands of his job, when in fact his lunch date issues from a different disposition. Her interpretation of the situation need not be motivated by personal desire, for she would have interpreted the behaviour of her friend's husband in exactly the same way.

Pears makes much of these "cold" cases of irrational belief-formation, but I do think that the central and most dominant cause of self-deception is that of a desire or a wish. It is no coincidence, I think, that self-deception is usually practised in situations that have a high emotive factor: in situations of love, self-esteem, maternal feelings, fear of failure, etc. In later discussions of "weak" and "hard" cases of self-deception I shall concentrate on desire-motivated cases, i.e. "hot" cases. In the following section which deals with the most common strategies of self-deception, I shall look at how these strategies aim at the desired belief, a belief backed by the agent's emotive involvement.

Pears suggests (1984, p.61) three different strategies employed in self-deception. Either self-deception operates directly on the contents of the mind in that it biases the processing of what is already in the mind, or it filters input into the mind, or there is the strategy operating through output, acting as if something were so in order to generate the belief that it is so, that is to act as if the desired belief were true. This last strategy is that of self-deception in the employ of akrasia, or what Pears terms self-deceptive akrasia. These three strategies do not necessarily operate separately, but they may also be employed as mutually reinforcing strategies. But more of this later on in the chapter. I have already made passing references to these various strategies of self-deception, but I now want to look at them in more detail. In Chapters 4-6 I shall be referring back to these strategies, when I discuss the "weak" and "hard" cases of self-deception which rely on either one or more of the following strategies for their success.

The first strategy, and the most common in cases of "weak" self-deception, is that of self-deception operating through input. In Chapter 1 I looked

at various ways in which one person can deceive another and showed that depriving another person of relevant evidence is sometimes a way of deceiving him. Using this as a base for self-deception, I can say that controlling the input of information into one's own mind counts as a strategy that may lead to self-deception. This strategy aims at preventing the evidence, which the agent suspects may be damaging to his favoured belief, from being registered in the mind and the key problem of internal irrationality is thereby avoided, provided the agent does not adhere to the second-order principle of the requirement of total evidence, as discussed in the section on akrasia in the previous chapter. By deliberately selecting evidence that supports the favoured belief or by wilfully avoiding suspected counter-evidence or even by highlighting the minimal evidence which will deflate the unwelcome belief, the self-deceiver carefully selects the input into his mind. Pears maintains that if there is a paradox in self-deception when it is done in this way through filtering, "it will be the paradox of irrational action, because it is a kind of akrasia to avoid maximizing relevant evidence and to go for unfair examples". (1984, p.63) This paradox, however, applies only to the self-deceiver who knows he has better reasons (already registered in the mind) for accepting the negation of the proposition he accepts and who knows he ought to accept that proposition for which there are better reasons. For the self-deceiver who filters reasons only for the acceptance of the proposition into his mind, and who avoids input of better reasons for accepting the negation of the proposition, the difficult problem of inner irrationality does not arise since "better" reasons have not had a chance to be registered in the mind and he can happily accept the proposition on the basis of only the selected reasons which are registered in his mind. But matters are not as simple as that. I shall have to deal with Sartre's point that in order to avoid the truth, the agent has to know very exactly just what that truth

is. However, I shall leave this problem aside for the time being, but shall return to it in Chapter 5.

Sartre's observation shows that evasion of relevant data implies purposefulness; it is not merely a random avoidance of information conducted in a haphazard way. The filtering of input or evasion of certain information implies avoiding something by using skill, cunning and strategem. The practitioner has a suspicion or intimation that something unpleasant will be uncovered if he were to exercise his attention, reasoning or information-gathering skills in a certain direction or on a certain topic. It is on the basis of this suspicion that he proceeds with avoidance tactics. The strategy operating through input filters incoming information "into the mind by avoiding looking for evidence where it seems likely that it will go against the favoured belief and by looking only where it seems likely that the evidence will support it." (Pears, 1984, p.63)

How will this strategy be employed by the girl who believes her lover to be faithful, despite strong counter-evidence? Her wish for the faithfulness of her lover causes her to wish to believe that he is faithful. She forms and accepts this welcome belief. However, he has come home very late on various occasions, makes strange mumbled telephone calls, and has been seen with another woman at a certain café. The girl suspects that there is a possibility that he is not faithful and suspects that she may find something unpleasant if she were to delve too deeply. She is not prepared to undertake an investigation into her lover's activities for this action carries with it the risk of discovering that he is not faithful. (It could, of course, also show that there is nothing suspicious about his activities, that the meetings are legitimate business meetings and that he is, in fact, quite faithful.) So whatever the outcome, she is not prepared to embark

on that first step. She thus avoids the café so that if there is a chance of her lover being there, she won't surprise him——the strong counter-evidence of his fidelity thus never arises for it is never allowed an opportunity to be registered in her mind. When he comes home late, she doesn't question him, in case he may say something she doesn't want to hear. Apart from avoiding likely damaging evidence she may also seek to promote supportive evidence for the favoured belief. She may seek out special occasions or do certain things which she knows makes him appreciative towards her. His appreciation of her culinary skills may then be seen by her as supportive evidence of his love and, therefore, his faithfulness. She is, thus, controlling the input into her mind and avoiding input that may lead her, as a competent rational being, to form and accept the unwelcome belief that he is not faithful. She avoids certain input because she wants to believe that her lover is faithful. By employing the strategy of selective input the self-deceiver either avoids the evidence for the unpleasant belief that not-p or stresses the evidence for the welcome belief that p without having to form and hold both beliefs, since to have a mere suspicion that not-p is not necessarily to have a strong belief that not-p. The girl is guilty of "weak" self-deception. Although her behaviour of not questioning her lover's dubious excuses, and of not confirming her suspicions may be judged irrational by us, it need not necessarily be internally irrational, or a genuine deviation from her own norms. However, if she holds the second-order principle, the principle of the requirement of total evidence, yet still refuses to acknowledge damaging evidence or deliberately seeks only evidence that supports her favoured belief, then she is guilty of inner inconsistency. In other words, if, as a rational competent being, she does hold this principle and yet flouts it by deliberately avoiding evidence for not-p, then we are faced with the paradox of internal irrationality.

The second strategy of self-deception operates through output. This brings us closer to self-deceptive akrasia, in which the agent acts as though the desired belief were true. The strategy generates the belief by a rather complex causal linkage, a mechanism by which the belief that one holds a belief makes itself true. Pears describes the workings of the mechanism in the following way:

> "Because beliefs normally monitor themselves accurately, it is an almost irresistible assumption that, if one believes that one holds a belief, that must be because one really does hold it and it is monitoring itself in the usual way ... The next thing that happens is that we rationalize the assumption that the belief is monitoring itself in the usual way by actually acquiring the belief. This objective rationalization is possible only when the rationalizing object is in the mind." (1984, p.59)

This strategy of self-deception reverses the usual order of things, because the person acts in order to produce the belief that would normally support the action. In Fingarette's account of self-deception (1969, ch.4), the strategy of behaving as if one already had the desired belief occupies the central place. To behave in such a way is to stimulate or fortify the belief on which it would be based in a rational structure. The complex causal linkage of the normal flow from belief and desire to action is reversed and the reversal is exploited by this particular strategy of self-deception. Davidson (1985, p.143) points out that when we say "Charles has a reason to believe that p", the meaning of "reason" is ambiguous; it can refer either to:

(1) evidence one has for the truth of a proposition (a cognitive reason) or

(2) it provides a motive for acting in such a way as to promote having a belief (an evaluative reason).

In the strategy which operates through output, the agent relies mainly on the second type of reason. In self-deception the wish that p were the case can easily give rise to the wish to believe that p, and this desire in turn can lead to thoughts and actions (reasons of the second kind) aimed at or resulting in obtaining reasons of the first kind, cognitive reasons for holding that belief. For example, I am frightened of the dark, but wish that I were not scared while walking along the gloomy forest path. I am motivated by the desire not to be scared, so I whistle in order to keep the fear at bay. I say, I can't be frightened, look at the carefree way in which I whistle. In other words, my action provides evidence (a cognitive reason) for the belief that I am not frightened. The action in the first place has been motivated by the desire (I acted on something already in the mind) and the action, in turn, provides evidence for the belief (I acted in order to put something in the mind). What must be distinguished is the two-fold role of the action: acting as a result of the desire to appear at ease, and acting in order to provide evidence for the belief that I am not scared. Furthermore, the action, motivated by the desire, can have as its aim either to suppress an unwelcome belief and to prevent it from arising (e.g. the girl who overburdens her already heavy workload so that there is no time to reflect on her lover's infidelity) or to foster a belief (e.g. the girl who constantly manoeuvres herself into her beloved's company and so deduces from the fact that they are always together that he must love her). As Davidson notes, there is nothing necessarily irrational about performing an intentional action that aims to relieve fear or distress or to arrange matters so that one comes to have a certain belief. If one were to arrange matters in such a way as to instill a certain belief in someone else, the action may be regarded as immoral —— especially if one knew or suspected the belief to be false —— but it is not necessarily an irrational

action. So too, Davidson maintains, when one does this to one's future self (as in trying to instill in myself the belief that I am not scared, even though my knees are shaking at the time) the action is not necessarily irrational. What does make it irrational is if one continues to cling to the fostered belief even though one continues to think that the evidence against the belief is better than the evidence in its favour. Of course, if the desire produced the belief without providing any evidence in favour of the belief (albeit "evidence" of one's calculated action), the belief is irrational.

> "One usually does something because one has a reason for doing it, but in this kind of case the agent does something in order to give himself what would have been a reason for doing it. The structure exploited by this strategy, though irrational, is a familiar feature of our lives. There is the 'sour grapes' reaction of those who miss something good and what some call the 'sweet lemons' reaction of those who get something bad. This is the territory of cognitive dissonance."
> (Pears, 1982, p.279)

The operation of the mechanism of this strategy can be compared to the similar operation of a sense-perception mechanism. "(P)eople often think that they can see something that is not there for them to see, because they have inferred that it must be there, and then in a certain sense they really do 'see' it." (Pears, 1984, p.59) In other words, people often think that because they have acted in a certain way, they must have had a belief which gave rise to the action in the first place. By thinking that they had a belief on which the action was based, they then come to acquire that belief.

Pears substantiates this view by referring to the experiments done by cognitive psychologists, in which students were asked to tell a lie, for

which they would receive payment. The assumption was that they were all opposed to lying, but the money was the motivating factor for akrasia. The assumption that they are all opposed to lying is an important one. It implies that the students strive to be moral. They, therefore, experience cognitive dissonance when their action of lying is in direct conflict with their accepted code of ethics. If they had no such code of ethics, the telling of the lie would not bother them in the first place, and there would thus be no need to embark on a self-deceptive project. The students who lied experienced cognitive dissonance after they had told the lie; they experienced two things which they found hard to accept together, i.e. the telling of the lie for money and the belief that lying is wrong. One way in which students tried to reduce the cognitive dissonance was that they altered their beliefs about the wrongness of telling the lie before they told it. Another possibility exists: the students told the lie against their own unbiased better judgement and altered their beliefs about the wrongness of lying only after they had told it. Cognitive psychologists term this approach to the lie as a "change of attitude". However, as Pears points out (1984, p.58), what strictly speaking occured was a change in factual belief, which produced a change in attitude towards the particular lie but not, of course, towards lying in general. The students rationalized after telling the lie by emphasizing the mitigating circumstances and said to themselves that the lie was, therefore, not a serious one. Apart from rationalizing, they also backdated their new belief and attributed it to themselves when they told the actual lie. By backdating this belief, the students in retrospect succeed in making for themselves the telling of the lie a "rational" act, rather than an akratic act, an irrational act against their better judgement. So not only did they have a new attitude to the particular lie, they also had a new belief about that attitude. The distinction that

Pears makes here is, of course, the distinction between holding a particular belief and believing that one is holding a particular belief. By referring to these experiments, Pears wants to stress the somewhat unfamiliar causal link in which action leads to belief. "It is normal for belief to lead to action, but it is possible for action to lead to belief." (Pears, 1984, p.60)

Two distinguishable stages in the "output strategy" or "recoil strategy", as Pears terms it, are noted:

(1) The self-deceiver acts and by acting makes himself believe that he has the belief that justifies his action, and next

(2) by believing that one holds a justifying belief, this latter belief becomes a permanent feature in his life.

Subsequent actions will, therefore, re-inforce the self-deception. How will the girl use this strategy to deceive herself? She may be making an error about herself by inferring that, because she is continuing the relationship, she must believe that he is faithful. As a rational being, she can make sense of her own behaviour only by representing it to herself as rational behaviour. She attributes to herself the belief that he is faithful; a belief that makes her action of continuing the relationship rational. Because she believes that she has the belief of his fidelity, this latter belief becomes a permanent feature in her life. Her continued relationship, of course, reinforces the belief and in a case like this the later admission of the falsehood of the justifying belief will be more painful the longer the self-deception continues. She, of course, did not hold the justifying belief at the time when the self-deception was being initiated, because what she actually did was to continue the relationship

in spite of the belief that he was not faithful. A paradox I noted in Chapter 1 is that self-deception is a vice peculiar to people who strive to be rational; to people who wish to mask the irrationality or their belief or action from themselves. Just as the students who wish to evaluate themselves as moral agents experience cognitive dissonance, and as a result embarked on a self-deceptive project, so the girl who wishes to evaluate herself as a rational being has to find a way in which her irrational action will be masked in a cloak of rationality. Her wish to rationalize her action generates the belief that would normally support that action. Whereas in the employment of the first strategy, self-deception operating through input, the girl makes a mistake in interpreting her lover's behaviour, in the case of the second strategy she makes a mistake in interpreting her own behaviour.

An additional illusion to this strategy is that when she achieves the belief that she believes p, it will seem to her as if she is monitoring it directly, whereas it is inferential. A difficulty arises, however: "It might be objected that people cannot really make mistakes about their own beliefs; they can lie when they report them, but they cannot make mistakes about them." (Pears, 1984, p.48) Pears points out that this is not necessarily so. One common mistake the girl has made is a mistake about her past belief——nobody can doubt the possibility of this kind of mistake being made about a belief of some time back. Secondly, she can make mistakes about her own intentions. Often people say sincerely that they intend to do something when in fact they do not, as shown clearly by their actions. So, one can make mistakes about one's own beliefs, ensuring the success of this strategy, and of self-deception in general, in everyday situations.

The third strategy operates directly on the contents of the mind. It biases the processing of what is already in the mind.

> "When the processing of information already in the mind is biased, there is seldom anything that the self-deceiver does in order to bias it. Sometimes, no doubt, he will avoid working out the implications of a belief, but usually the biasing is the direct effect of the wish and there is nothing that could be regarded as a plan." (Pears, 1984, p.61-2)

The normal processing of the contents of the mind follows the following sequence: achievement, belief in achievement and satisfaction. The wish, however, exploits this sequence when actual achievement is too difficult to bring about. As I discussed in the section of a wish as a cause of self-deception, the biased sequence is as follows: the wish for achievement, the wish to believe that one has achieved, the belief in achievement and satisfaction. The wish aims at satisfaction and when actual achievement is not possible, the wish aims at the belief in achievement that will satisfy. The belief, therefore, is the "messenger with good news". The self-deceiver does not formulate a plan, but relies on the "boyancy" of the wish to generate the belief that will satisfy. The "discreet operation" of the wish to believe that p keeps the self-deceiver ignorant of his own self-deception. However, if there is some plan which the self-deceiver formulates, it will have to be ascribed to a sub-system within the person —— a sub-system that will note the weaknesses in "the rest" of the person and devise strategies to exploit them. This possibility will be dealt with in the Theory of Systems in Chapters 5 and 6.

A last note I want to add is that strategies are not always separately employed. Often two or more strategies are used to reinforce the irrational belief or action. The girl might be using the strategy of continuing to

act as if she believed him faithful in order to generate or fortify the justifying belief, but her action, on the other hand, may show selected avoidance of certain "danger areas", such as the café or refraining from questioning him about his whereabouts.

The self-deceiver employs the above strategies in order to mask some unpleasantness or to generate some satisfaction by holding a welcome (albeit irrational) belief or by avoiding an unwelcome belief. This is one striking difference between self-deception and other-deception. An other-deceiver can deceive someone else for no ulterior motive whereas the self-deceiver always does have an ulterior motive. "It is .... not clear what could be meant by or what justification there could be for, speaking of somebody as deceiving himself if it were at the same time contended that what he was said to be deceiving himself about was a matter of total indifference to him, in no way related to his wants, fears, hopes and so forth: could we, e.g., intelligibly talk about 'disinterested' or 'gratuitous' self-deception?" (Gardiner, 1970, p.242). Pears too feels that it is impossible for a person to deceive himself purely for the sake of deceiving himself. The reason for this is "the ineradicability of the desire for the truth of one's own beliefs", whereas "the desire to impart truth to others is not ineradicable." (Pears, 1984, p.42). Although the truth of a person's beliefs is not always the ultimate goal, it does, however, retain some attraction when he is forming the belief, especially since the self-deceiver with which I am concerned is a rationally able and competent person.

Moreover, the strategy of self-deception which operates through output, acting as though the belief were true in order to make himself believe that

he has the belief, seems to have no exact parallel in the case of other-deception. The other-deceiver can hardly perform his victim's actions for him, and even if the other-deceiver acts as if the belief that he wants his victim to hold were true, it is the action of the deceiver that is rationalized and not the action of the victim.(3) Pears, however, feels (1984, p.62) that the parallelism between other-deception and self-deception need not be so exact in order to justify the classification of the strategy which operates through output as a method of self-deception.

Kipp notes a further alleged difference between self-deception and other-deception: "Normally, when attempted other-deception has obviously failed to dupe its intended victim, the deceiver resigns himself to admitting that the game is up, and abandons his project; but when attempted self-deception has obviously failed in a similar way, the deceiver characteristically intensifies his pretence, and persists in the game with desperate grotesqueness." (1980, p.313). I would certainly agree with Kipp that this does happen in most cases of self-deception and other-deception, but it does not necessarily happen in <u>all</u> cases. For example, the convicted criminal may persist in proclaiming his innocence even though he himself knows that he is guilty and his guilt has been proved beyond doubt to the judge. "The game is up", and yet he clings to the irrational belief that if he persists for long enough, the judge will believe him. Or the self-deceiving terminally ill cancer patient may, after being confronted with further conclusive proof, accept that he is dying of cancer. Of course, when "the game is up" there is no longer deception, other or self, but the main point I want to stress is that Kipp is not justified in regarding the above as a definite difference between self-deception and other-deception.

Kipp, however, does note a more striking difference between self-deception and other-deception: that of primary and secondary deception. In order to illustrate this point I shall return to the general definition of other-deception as offered in Chapter 1 :

(1) A gets B to believe that p

(2) A suspects/knows that p is false.

A, the deceiver, knows the truth, yet intentionally gets B to believe the opposite. We have here primary deception only: the deception of B. A need not deceive himself about anything. He knows the truth; he knows that he intends to deceive B; he knows that he tells B something that is false. There is no mental conflict or inner irrationality here. However, when A deceives A, he knows the truth and yet persuades himself of something which he knows is false (the epistemological paradox as discussed in Chapter 1). But, Kipp maintains, there is a further requirement for successful self-deception: the self-deceiver must not only reconcile simultaneously held, conflicting beliefs, i.e. aBp+aB-p (primary deception) he must also deceive himself about the belief that he holds conflicting beliefs, i.e. aB(aBp+aB-p) (secondary deception). Therefore, self-deception seems to require primary deception about some unwelcome belief, as well as secondary deception about the unwelcome belief that he holds two incompatible beliefs. Again, I disagree with Kipp that this is a difference between all conceivable forms of self-deception and other-deception. In "weak" forms of self-deception successful deception can take place without necessarily forming the opposite belief. Successful self-deception may be attained by merely avoiding the unwelcome belief without necessarily holding an opposite belief.(4) The question of secondary deception does, therefore, not arise. It may be argued, of course, that the self-deceiver

must, in a way, be unaware of the fact that he is deceiving himself, if his deception is to be successful. He is, therefore, deceiving himself about deceiving himself, which is a form of secondary deception. This may well be so, but the point I am making is that Kipp's formulation of secondary deception, i.e. the deception about the belief that one holds conflicting beliefs, is not applicable in cases of "weak" self-deception. However, in cases of "hard" self-deception, the agent does hold conflicting beliefs and for this kind of self-deception to be successful, secondary deception does seem to be a vital requirement.

The above differences between self-deception and other-deception raise the question of whether the model of other-deception can then be used as a basis for self-deception. I do not, however, regard these small differences as constituting a threat to the justifiability of using the other-deception model. After all, as I mentioned before in Chapter 1, a step-by-step analogy is not possible, for the logic of other-deception does not march exactly in step with that of self-deception. But what I am aiming for is to examine "analogies and similarities with cases of deception proper that are sufficient to make the reflexive extension of the concept appear, within limits, reasonably appropriate. But the instances themselves will form a variegated spectrum, and the analogies can in any event never be more than partial ones". (Gardiner, 1970, p.243).

In Chapter 1 I looked at the various problems and paradoxes that arise when self-deception is modelled on other-deception. For fear of repeating myself, I shall highlight the four main paradoxes as noted by Pears (1974, p.98). So, before looking at ways in which to explain how self-deception is possible, I am going to note only the main problems which all

credible interpretations of self-deception must address.

(1) I believe p, but at the same time I believe that not-p. It seems as though the self-deceiver must believe in a straightforward contradiction. As Szabados points out: "The self-deceiver, as deceiver, must be aware of the truth; as deceived must be unaware of the truth". (1974, p.52) So if A deceives himself, "it seems that he must be motivated by his belief that p so that he is aware that p and this awareness leads him to get himself to believe that not-p. The belief that p and the belief that not-p are not occuring at different times in disparate areas of his intellectual concern but are intimate parts of the same endeavour." (Brooks, 1986, p.245-6). An assumption of this paradox is that the original belief should persist right up to the end of the process of deception which has as its goal the installing of the opposite belief, as well as assuming that it happens within the same mind (to exclude cases of schizophrenia or, rather, multiple personality). As Sartre states the paradox: ".... I must know the truth very exactly in order to conceal it more carefully — and this not at two different moments, which at a pinch would allow us to re-establish a semblance of duality — but in the unitary structure of a single project." (Sartre, 1958, p.49)

This first paradox is really that which has attracted the bulk of philosophical interest. It encompasses most of the problems encountered in the discussion in Chapter 1 on the epistemological paradox.

(2) The second paradox moves onto the problem of the psychological paradox as discussed in Chapter 1: I cannot intend to screen something from consciousness if the intention is motivated and guided by the

continuing awareness of that very thing. In other words, I cannot intend to do what I know I cannot do. If it is impossible for me to believe in the conjunction of two contradictory propositions, then I cannot intend to believe in such a conjunction. Since I know that the conjunction of incompatible beliefs is impossible, I cannot plan to bring about this impossible combination.

(3) The third paradox follows from the second. Since it is impossible for me to accept the conjunction of two contradictory beliefs, I must somehow keep them apart in my mind. The theory of Systems approaches the problem by allocating the two contradictory beliefs to two separate systems (in early Freudian terms these would constitute the conscious and the unconscious). This solution, however, gives rise to the third paradox: I cannot divorce my belief that not-p from the rest of my thoughts. In order for the plan to be motivated and guided, an awareness of the initial belief is needed. If I do not have the belief not-p (either I have failed to form it or I have forgotten it), I cannot take preventative steps. Being self-deceived implies an indentification of its cause. I believe p because I believe not-p. The unwelcome belief, not-p, sustains the favoured belief, p. But then we have the paradoxical situation where the unwelcome belief is a causal condition of a belief that contradicts it. It is because I am aware of the unwelcome belief that the need for self-deception about the favoured belief arises.

(4) The fourth paradox is an extension of the third — not only the belief, but the whole plan must be concealed from my thoughts and beliefs. If the plan itself is incoherent and impossible to put into practice, it cannot be made possible merely by being screened off from consciousness so that there is no identification with its cause.

I shall take these four paradoxes to cover the main problem areas in self-deception. Although, as I have shown in Chapter 1, there are numerous other paradoxes, I believe that any coherent theory of self-deception will have to deal adequately with these four paradoxes at some stage before it can be accepted as a plausible theory. This then is the formulation of the problem to which the rest of this study will be addressed.

The paradoxes and their solutions can be approached in two traditional ways: one can deny that literal self-deception exists and that the paradoxes, therefore, do not arise; or one can assert that there is such a thing as literal self-deception with all its accompanying paradoxes. If one were to hold that literal self-deception is impossible, then those issues which we label "self-deception" are merely metaphors which denote the employment of various psychological devices which enable the "self-deceiver" to hold on to his favoured belief (or to avoid an unwelcome one). If self-deception is merely a metaphor, then those phenomena which resemble a type of deceiving of oneself are falsely called self-deception. Often in these cases the question of the belief that p and the belief that not-p does not arise, or, if there is anything resembling the beliefs that p or not-p, these are beliefs that in no way entail certainty about either belief—more appropriately, these are "subjective inclinations" to seek, favour or emphasize available evidence that either underscores the favoured belief or undermines the unwelcome belief. On the other hand, one can boldly assert that literal self-deception is possible and that the resultant paradoxes can be explained. Now it seems as though only this last view uses the notion of other-deception to explain self-deception and that the former view of "weak" self-deception rejects the model of other-deception.

When I refer to the two different solutions to the four paradoxes, I do not imply that by accepting one solution, the other must be rejected. In fact, much of the debate on the question of self-deception seems to be locked in this either-or stalemate.(5) I want to show that instead of the acceptance of one solution excluding the other, the two solutions are complementary to each other: the first solution is more applicable to cases of self-deception in which there is a certain amount of latitude, and the solution of the theory of systems is more applicable in severe cases of self-deception. I have already referred to this "sliding scale" of severity and will look at the solutions applicable to the different degrees of self-deception.

* * * * * *

In this chapter I have examined the various causes of self-deception and although forgetting, misperception and cold perversions of reasons do constitute causes of self-deception, I shall concentrate in the rest of the study on the rebellious wish as the main cause of self-deception. Although I regard desire-motivated self-deception as the central form of self-deception, I referred to examples with other causes to show the different forms of self-deception operating on the periphery of irrational belief-formation. Pears especially is interested in these, what he terms "cold", cases of self-deception, but I shall concentrate on the "hot" cases which, to my mind, are the central ones.

The "tools" or strategies which the self-deceiver employs also vary, and

either one or a combination of strategies may be employed by the self-deceiver in his quest for the favoured belief. Again, the strategy that operates through output does not constitute the central strategy of self-deception, although it is often used to re-inforce an already held irrational belief. The strategy which operates through input is the central strategy employed by the self-deceiver in "weak" cases of self-deception, whereas the strategy which works on what is already in the mind seems to constitute the main strategy in "hard" cases of self-deception. Of course, these strategies are not always used in isolation, but are often used in various combinations to contribute to the success of the self-deceptive project.

However, the strategies employed by the self-deceiver are not always the same as those used by the other-deceiver. Other differences between self-deception and other-deception seemed to threaten the plausibility of using the model of other-deception for self-deception. However, since these differences are minimal, I want to retain the other-deception model which clarifies, more than complicates, the notion of self-deception. By using the model of other-deception, four main paradoxes of self-deception arise and the rest of this study will concentrate on the various approaches to these paradoxes.

Notes

1. The results and implications of the experiments conducted by Festinger's school of cognitive dissonance will be discussed more fully in Chapter 4.

2. J. Penelhum (1964) and Canfield and Gustavson (1962).

3. It may be argued that post-hypnotic suggestion is a counter example: the victim acts on the suggestion and then rationalizes his action and not that of the hypnotist. However, as noted in Chapter 2, hypnotic suggestions do not fall within the scope of intentional deception——what the victim must rationalize is his free and intentional action.

4. C.F. Bach (1981) for whom self-deception entails avoiding the unpleasant thought p, without the agent having to form the opposite belief not-p.

5. E.g. Kipp (1980) who postulates two opposing teams, i.e. the minor stream with dry reasonableness on its side, which tends to deny that self-deception is possible, and the major stream with bold provocativeness on its side, which claims to know that self-deception exists. "If self-deception is something very similar to other-deception, it seems to imply paradoxical states .... If, on the other hand, self-deception is something very different from other-deception, it seems to run the risk of not being 'deception' at all ... Among theorists who wish to avoid the paradoxes without having recourse to question-begging concepts like unconscious believing, half-believing, unnoticed believing, or multi-selved believing, it is usual to renounce any close analogy between other-deception and the phenomenon we call 'self-deception'." (p.305).

CHAPTER 4

THEORIES OF "WEAK" SELF-DECEPTION

I concluded the last chapter with the view that an "either-or" approach runs the risk of excluding a vast range of cases of self-deception. On the one hand, if self-deception is totally divorced from the notion of other-deception in order to circumvent and deflate the paradoxes, self-deception seems to run the risk of not being deception at all, and it will hold little interest as a speculative problem. If, on the other hand, self-deception is interpreted strictly according to the model of other-deception, it seems to imply paradoxical states like knowing and not knowing, i.e. literal self-deception. These two extremes occupy opposing ends of the scale: on the "weakest" end the word "self-deception" refers metaphorically to a cluster of phenomena that we falsely call self-deception, and which are really types of wishful thinking. On the "hardest" end the word "self-deception" refers literally to the paradoxical state of a person not only holding consciously two contradictory beliefs (i.e. $aB\underline{p}+aB-\underline{p}$) but requires that the self-deceiver must also consciously conjoin two contradictory beliefs (i.e. $aB(\underline{p}+-\underline{p})$). This, however, seems an impossible feat for any sane person.

In this chapter I want to show that "self-deception" is not merely a metaphor for irrational belief-formation, but that a person can deceive himself without necessarily being confronted by the insolvable paradoxes. According to my sliding scale of other-deception, a genuine, full-blooded case of self-deception does not necessarily mean a lie-to-oneself (i.e. based on case (6) of other-deception). Proper self-deception can also be based on cases (1) - (5) of other-deception and self-deception can thus entail the

avoidance of certain suspected counter-evidence, the finding of pseudo-reasons for a favoured belief, etc. In Chapter 1 I referred to Pears' sliding scale of cases of self-deception, and in this chapter I shall concentrate mainly on the first two cases:

1. Balanced evidence for p and not-p, accept p.
2. Inductive evidence for not-p, accept p.
3. Deductive evidence for not-p, accept p.
4. Not-p, accept p.

How interpretations of cases (1) and (2) confront (or, rather avoid) the paradoxes will form the main part of this chapter, and cases (3) and (4) will be dealt with later.

First of all, I shall examine how cases (1) and (2) are analogous to cases of other-deception, since these serve as my model for cases of self-deception. Cases (1) and (2) of self-deception are analogous to cases (1) - (5) of other-deception as set out in Chapter 1, i.e. from unintentional other-deception to the manipulation and exploitation of available evidence in order to instill a false belief in the other person.

> "Interpersonal deceivers evade acknowledging to others something they know, believe, suspect, feel, and so on. Often they engage in pretence, withhold their emotions, prevent others from having explicit consciousness about something, keep others ignorant, or persuade them to hold false beliefs. There are analogies here with self-pretence, emotional detachment, systematic ignoring, willful ignorance, and rationalization."
> (Martin, 1986, p.20)

When A withholds evidence from B so that B is led to form a false belief that p, A has deceived B, i.e. has intentionally caused B to form a false belief. Just so, if A avoids evidence he suspects may prove to be

damaging to his favoured belief and, through deliberately not collecting all the facts, A holds a false belief that p, we must accuse A of self-deception. He has intentionally caused himself to hold a false belief. It certainly is a "weak" case of self-deception, but since there are also "weak" cases of other-deception, we are entitled to attach the label of self-deception to the former. Self-deception in only its extreme form is a "lie to oneself" (case (4) of self-deception and case (6) of other-deception), but there are many different ways of deceiving others than lying to them, just as there are different forms of self-deception which do not entail this limiting case of lying to oneself. It is for this reason that the "either-or" choice is a postulation of an erroneous approach. Within the different forms of self-deception there are no definite exclusions, but rather graduations of degrees of irrationality.

Before looking at the specific forms of self-deception as mentioned above, it may be appropriate to look at the general shift in emphasis in "weak" self-deception in order to get around the paradoxes. In "weak" self-deception, the inner tension experienced need not always be a matter of beliefs. According to Martin (1986, p.27) inner division can include ambivalent emotions or attitudes, conflicting desires, self-contradictory inclinations to believe, etc.(1). Whereas "hard" self-deception centres on false beliefs, "weak" self-deception includes a wider variety of forms of self-deception. Martin and Fingarette hold that intentional self-deception can also involve the purposeful or deliberate evasion of full acknowledgement of something to oneself. The manoeuvres of the self-deceiver in this case are not centred on belief formation. This view will be examined a little later on.

As with other-deception, self-deception typically involves concealing a

truth or one's view of the truth, but the concealment can take many forms. It may seem that the very notion of truth will steer us directly back to the epistemological paradox, involving true and false beliefs, as well as the psychological paradox in which the motive for self-deception is the rational person's desire to be rational and to hide his irrationality from himself by embarking on an irrational project. However, this paradox is based on the notion that beliefs are truth centred in that "they involve efforts to acquire knowledge. Yet none of us is so completely dedicated to truth that we do not have competing intentions and motives as we form our beliefs. We seek the truth; we also seek to adopt beliefs that make us happy, support our self-esteem, provide a comfortable worldview, and align with our basic loves and committments. Self-deception is often a special case of forming and holding beliefs on the basis of mixed concerns." (Martin, 1986, p.26). This, of course, does not mean that the self-deceiver is totally blind to the influence of his bias or that he has no notion of the truth which he is trying to evade or hide. A self-deceiver usually forms and holds a favoured belief by intentionally disregarding counter-evidence, but the self-deceiver can also intentionally ignore the biasing influence of his wish, or intentionally avoid making enquiries which he suspects, or even knows, are appropriate. Although the self-deceiver in this case is aware of evading the truth, it does not imply an impossible paradoxical situation, since he is not exactly sure what the truth is. Therefore, rather than the problematic conflict between true and false belief, what we have in this case amounts to an <u>absence</u> of true beliefs and, hence, an absence of contradiction. However, the characteristic of mental conflict is still present, although not as acutely felt as in more severe cases of self-deception. There is no contradiction of beliefs, but there is inner tension between holding the favoured belief and constantly avoiding evidence that may undermine that belief. Although he does not

know that not-p, he avoids not-p. In saying this, we need not describe the self-deceiver as "intending to believe what he knows is false." It is enough to describe him, in this "weak" case, as intending to disregard evidence, to emphasize other evidence, to avoid enquiries, to evade self-critical scrutiny of possible biases, etc.

The emphasis in "weak" self-deception, therefore, shifts to a watering-down of irrationality. The paradox is considerably weakened if the view is held that people are not always dedicated to the truth, but that their desires for a certain state of affairs takes precedence over the actual truth. Furthermore, if there is uncertainty about just what the truth is, it can lead to a suspicion rather than a belief about the actual state of affairs, or even, as some philosophers claim, entail an absence of true beliefs. Other philosophers again evade the paradox by interpreting self-deception not at all in terms of beliefs, but rather in terms of "refusal to spell-out one's engagement in the world", "self-acknowledgement" or "attempts at self-modification"(2). This chapter will deal with desire-goaded "weak" forms of self-deception (including such forms of self-deception as evasion, jamming, failure to focus, mental distance, biased thinking, rôle-playing, rationalization, etc.) and how the interpretations of these aim at overcoming the paradoxes, especially the epistemological paradox which seems to underlie the others.

Since the first case of other-deception is unintentional deception, I shall first of all look at cases of unintentional self-deception and look at various criteria which distinguish these from cases of intentional self-deception. Obviously, unintentional self-deception does not raise the psychological paradox of intending to deceive oneself, but I shall show how unintentional self-deception does not raise the problem of the epistemological

paradox either. Mele's approach (1983) to resolving the epistemological paradox is to reject the assumption that knowledge is a prerequisite for all forms of deception.(3) Mele, of course, does not deny that knowledge or true belief is a prerequisite for intentional deception, but he does not agree that it is a prerequisite for all deception. "(S)urely A may induce a false belief that p in B, and thus deceive him in this sense, without knowing, or even believing, that p is false?" (Mele, 1983, p.366). Mele adds that A may indeed believe p to be true, and he may have intended to communicate it to B by telling him that p. A has deceived B, but it is not an intentional deception, i.e. A did not deliberately set out to deceive B. However, Mele's observation that a true belief may be a prerequisite for intentional deception raises an interesting point. How then are we to classify the case in which A does intend to deceive B by telling him that p, but unbeknownst to A p is, in fact, true. Thus, if the deception of B is successful, B will come to believe truly. This, I think, is a case of intentional deception and although it does not involve factual knowledge of the truth of p, the case still involves knowledge and "true" belief of some sort—knowledge which A thinks is factual, i.e. he sincerely thinks that not-p is true, and he treats not-p as a "factually true" belief. So Mele's statement is a little ambiguous since we can have a case of intentional deception that does not include a true belief. So it seems rather that what the agent sincerely thinks is a true belief and a deliberate attempt to deceive with respect to this belief are prerequisites for intentional deception but not for all deception. The distinction may be presented as follows:

1. A believes falsely that p — A communicates p to B — B believes falsely that p.

   This is a case of unintentional deception in which A caused B to be in error with respect to p.

2. A believes falsely that p ——A communicates not-p to B ——B believes truly that not-p.

   This is a case of intentional deception in which A deliberately gets B to believe the opposite of what A knows (or, rather accepts) to be true.

Mele's interpretation of "deception" refers to the loose definition of "deceive" i.e. to cause someone to be in error with respect to p. A's action which causes B to believe falsely that p is not a case of intentional deceiving, but it is a case of deceiving nonetheless.

Mele labours this last point for he aims to show that the "vast majority of cases of self-deception are not cases of intentional deceiving." (1983, p.366). He is quick, however, to note that the suggestion then that self-deception should be modelled after unintentional other-deception runs the risk of oversimplifying matters. Unintentional other-deception may be quite accidental, whereas self-deception seems to be motivated by the agent's wishes. Self-deception, therefore, is not accidental but, he stresses, "the non-accidentality of self-deception does not imply that the person must intentionally deceive himself. To be sure, self-deceivers often do engage in intentional behaviour with the result that they become deceived with respect to p; but ... they rarely act with the intention of deceiving themselves." (Mele, 1983, p.367). What does not happen, therefore, in the typical case of the girl who persuades herself that her lover is faithful is that at one time she had the explicit intention of "getting myself to believe what I know or believe to be false." Mele's point is that the self-deceiver does not aim at deceiving himself; he has no intention of deceiving himself. He may, of course, engage in intentional behaviour to find pseudo-evidence to support his favoured belief, but there is no explicit intention for deception as such.

I think Mele is quite correct in pointing out that the self-deceiver rarely has the explicit intention to deceive himself, but Mele is guilty of failing to distinguish between primary and secondary deception. Primary deception entails deception about p, whereas secondary deception entails deception about the process of deception itself. Self-deceivers typically use self-deceptive strategies to evade or hide from themselves the fact that they are using self-deceptive strategies, in other words, self-deceivers deceive themselves (secondary deception) about their self-deception (primary deception). In doing so they may use the same tactics at a second order level with respect to the primary use of the tactics. The target of primary deception is the main thing the person is deceived about, whereas the target of secondary deception is any related thing the person deceives himself about as a means of evading or falsifying the primary target. The self-deceiver thus has the intention to believe that p and his actions are guided by this intention (e.g. to evade certain counter-evidence, etc.), but he need not necessarily have the additional intention to deceive himself about his own self-deception. "But there can also be further motives for the secondary evasion which need to be concealed. In particular, it can be less than flattering to recognize that we have been deceiving ourselves and to admit to the reasons we have done so." (Martin, 1986, p.18). I do agree with Mele that few self-deceivers have an explicit intention about deceiving themselves about their own deception, but it is wrong to label this case then as unintentional self-deception. Most self-deceivers engage in intentional behaviour, but this behaviour is aimed at avoiding not-p and at forming the belief that p, rather than the self-deceptive process itself. The deliberateness with which the self-deceiver goes about his active search or willful avoidance will make no sense if he does not have the intention to seek out or avoid telling evidence.

However, if we employ the loose sense of "deceive" it is easy to see how one can be the cause of one's being in a state of error with regard to p, be it either through negligence, fatigue, impetuosity, misperception, an inability to think things through, etc. The problem, however, arises that these forms of unintentional self-deception are not easily distinguishable from cases of pure mistakeness or incompetence. For example, if the aforementioned girl has no intention of deceiving herself about her lover's faithfulness, but she does believe in his fidelity nevertheless, the irrational belief may be purely due to intellectual incompetence. I am sure that Mele is aware of these obvious objections that can be raised and he is, therefore, quick to note that intentional deception is "admittedly, the central case of deception", but not all cases of deception are cases of intentional deception, he adds. I do grant that, but I disagree with his view that unintentional self-deception forms the vast majority of cases of self-deception. Certainly, there are few cases of intentional secondary deception, but the majority of cases are of intentional primary deception. As I have shown in Chapter 1, unintentional other-deception does exist, as does unintentional self-deception, but these constitute a small minority of cases on the periphery of the kind of deception that is of philosophical interest.

Before going on to discuss the central case of self-deception, viz. that of intentional self-deception in which the agent engages in intentional action in forming a belief or finding or avoiding evidence for a certain belief, I want to note a distinction made by Palmer and Champlin (1977) that not all cases of deceiving oneself are cases of self-deception. In other words, not all self-reflexive deception is intentional. They hold that to deceive someone can be to do no more than to get him to mistake appearance for reality. Palmer offers the example of an inexperienced gardener who

deceives himself but is no self-deceiver. After consulting his newly-acquired gardening books, the gardener digs up all his snapdragon seedlings in the mistaken belief that he has positively identified them as weeds. The essential ingredient which distinguishes deceiving oneself (that is, unintentional self-deception) from self-deception (that is, intentional self-deception) is, according to Palmer and Champlin, "dishonesty with oneself." To substantiate this claim, Champlin offers an example of intentional other-deception which may lead to unintentional self-deception: A camouflage expert hides a gun so skillfully that not only does he fool the others into thinking it a clump of bushes (he, of course, intended to deceive the others), but he is even taken in by his own handiwork, (he unintentionally deceived himself). But it would not be natural to call this a case of self-deception, according to Champlin, since the camouflage expert is not being dishonest with himself when he was deceived by his own handiwork. Champlin's condition for dishonesty with oneself in a case of intentional self-deception steers us into interpreting self-deception as a moral notion. He compares the difference between intentional and unintentional self-deception to the difference between murdering and killing. "Murder is something you commit; ... self-deception is something you are guilty of." (1977, p.291). Dishonesty with oneself certainly does seem to be an ingredient in intentional self-deception, but I feel the distinction is certainly not the most important one. Champlin appeals to dishonesty with oneself on which he bases his view of self-deception, but the concepts he uses are obscure. Even if dishonesty with oneself entails an "interior dialogue" and self-deception is, therefore, a kind of "lying to oneself", it doesn't explain just how self-deception works or how it solves the paradoxes. Mele (1983, p.375), in fact, offers an example of a person lying to himself and who is, thus, being dishonest with himself, but the case still does not seem to be typical of self-deception proper. The example is of a

stuntman who prepares himself for a feat which he knows to be exceedingly dangerous. He may tell himself that there is nothing to fear, with the hope or intention of inducing in himself the belief that there is nothing to fear, even if it is only for the crucial moment.

The above interpretations of unintentional self-deception have all relied on a loose definition of "deceive" (i.e. "to get oneself to mistake appearance for reality" or "to cause oneself to be in error with respect to p") but the meaning of the word has a stricter application when used in intentional self-deception, i.e. "the action of deceiving oneself". What then is the distinguishing ingredient between the loose definition and the stricter one, or between unintentional deception and intentional deception? What distinguishes unintentional self-deception from intentional self-deception then seems to be an absence of knowledge of the true belief, the absence of the deliberate attempt to deceive oneself either at a first or second-order level, and the absence of deliberate dishonesty with oneself. It is clear that these three criteria for unintentional self-deception evade the main paradoxes. Firstly, the epistemological paradox is avoided since in unintentional self-deception there is only one belief——one in which the agent is in error——but one belief, nevertheless. The opposite belief, be it that the lover is in fact unfaithful, or that those seedlings are in fact snapdragons, is not registered in the mind. In unintentional self-deception we, therefore, do not have to deal with the epistemological paradox of trying to explain how the self-deceiver can believe p and believe not-p at the same time. The second criterion cuts out the psychological paradox because in unintentional self-deception the question of intending to deceive oneself is not raised. The third criterion removes the emergence of the ethical paradox in that the self-deceiver is wholly victim. I agree with Foss (with my modifications in parentheses) that intentional

"self-deception, like deception of another, has knowledge as a pre-requisite: one deceives himself that p ... knowing (or believing) p to be false." (1980, p.242). One cannot knowingly deceive unintentionally. According to Foss, "one must know one is deceiving, and that requires knowledge of what is really the case (or knowledge of what one thinks is the case)" (p.242). Whereas unintentional self-deception does not invoke the epistemological paradox of p and not-p, it seems as though intentional self-deception will have to confront it, i.e. knowledge of the false belief and knowledge of what is really the case.

I shall now look at various accounts of intentional self-deception which deal with the paradoxes by either watering them down (e.g. the agent is not absolutely certain that p and that not-p) or evading them altogether (e.g. self-deception does not necessarily entail the simultaneous holding of two contrary beliefs). One way of evading the paradox of p and not-p in a case of intentional self-deception is to appeal to the lapse of time. When A lies successfully to B, it doesn't necessarily mean that B immediately believes what A tells him. B may just think about the plausibility of what he has heard from A, without as yet accepting it as true and thus believing it. Similarly, when the stuntman forms the belief that there is nothing to fear, he tells the lie with knowledge that fear is warranted. This does not, however, necessarily imply that he believes that p and believes that not-p at the same time. Even though he tells himself that there is nothing to fear he may not form the corresponding belief until some time after he has told himself this. "The more I keep telling myself, the more chance I have of coming to believe it" ——this may even be a conscious strategy employed by the stuntman. He may come to abandon the true belief in the meantime. Demos offers (1960, p.591) the example of a young unattractive and lonely man who is only too aware of his inadequacies.

He, however, makes up for these by thinking of himself as a lady-killer. With time, he even comes to believe this and frequently convinces himself and others of his numerous romantic adventures. So, in lying to himself, he has changed his initial belief that caused distress, to one that now gives him pleasure. This points to the emphasis in "weak" self-deception in which there is an absence of the true belief and, therefore, an absence of contradiction. In order to generate a paradox in cases of successful reflexive lying, the liar must believe the lie immediately. If he believes the lie at the moment when he lies to himself, we can no longer appeal to the passage of time during which the true belief is eventually abandoned. Whether this case of "hard" self-deception is ever instantiated will be examined in the next two chapters.

I have mentioned earlier on that in cases of "weak" self-deception, the self-deceiver often deliberately evades some truth or certain issues or evidence. I shall now look at evasion as a form of self-deception, but it is not easy to see how deliberate evasion can solve the epistemological paradox, especially if we add the Sartrean stipulation that in order to evade the truth, the self-deceiver must know very exactly just what it is that should be avoided. To deceive oneself is typically to evade the truth or what one would view as truth if one were to face an issue squarely. One way around the problem posed by Sartre is to deny that the self-deceiver must <u>know</u> very exactly what he is avoiding. This approach relies on a re-description of what "know" entails when the agent must "know" that not-<u>p</u> but believes that <u>p</u>. There are two different re-descriptions of "know". First of all, know need not imply certainty. According to this description of "know", the self-deceiver may have a mere suspicion that some evidence on closer inspection <u>may</u> be damaging. I have already referred to this account in Chapter 2 when I examined the strategy that works through

selective input in the mind, and an account to which I shall return later in the chapter when I discuss the self-deceiver's appeal to latitude. The second re-description of "know" is based on the notion of awareness. Martin suggests that the self-deceiver need not know the truth in the sense that he is constantly aware of it. "Acts of deceiving oneself consist of the actions involved in forming and sustaining projects of evasion. A project of evasion can be carried out using an assortment of strategies and patterns of behaviour, but the project and its unifying intention do not involve continuous mental activity." (Martin, 1986, p.13). Martin compares this project of evasion to the project of avoiding someone's company: one does not constantly think about avoiding that person, but one employs a variety of dodging tactics on different occasions, like avoiding going to parties at which he is likely to be, or glancing away when he enters the room, or not responding to his conversation or conversation about him, etc. These tactics do not require constant awareness of that person, but when one is confronted by that person or by an aspect related to him, one simply turns one's attention away.

Bach also appeals to the notion of evasion to describe self-deception but, unlike Martin, he tries to free self-deception from paradox by arguing that self-deception involves a chasm between what one believes and what one thinks, and that self-deception should be interpreted in terms of action, rather than belief. "Self-deception is not essentially a matter of belief at all." (Bach, 1980, p.353-4). He argues that the self-deceiver may have the hidden belief that not-p without actively entertaining the thought that not-p. Therefore, if the thought that not-p never arises or crosses his mind, he will have no reason to deceive himself about the truth. The self-deceiver desires that p while having the belief that not-p (albeit not consciously entertaining the belief) and what he does

is "to avoid the sustained or recurrent thought that not-p".(4) However, before going on to the way in which the self-deceiver evades the sustained or recurrent thought that not-p, it is necessary to look at the distinction Bach makes between belief and thought, and on which he relies in order for his definition to hold.

According to him, beliefs are states, whereas thoughts are occurences. He notes that we have countless beliefs which we are not currently holding in the focus of our attention. We even have beliefs whose content we have never had in mind until now, e.g. that kangaroos are bigger than cockatoos. "It is enough that beliefs not be confused with occurences of the corresponding thoughts, i.e. the belief that not-p with the thought that not-p." (1980, p.355). This distinction allows for the fact that a person need not have thought everything he believes as well as the fact that his thoughts do not invariably correspond to his beliefs. Of course, Bach is not denying the obvious fact that usually what a person thinks, and thinks he believes, is in fact what he believes. What Bach is insisting on is that we cannot assume a person always believes what he thinks he believes, for then we would be ruling out the possibility of error about one's beliefs. He insists that the possibility must be allowed of thinking that not-p without believing that not-p. For example, a person is presented with persuasive arguments for and against p. When he hears the first, he thinks that p, but on hearing the second, he is more impressed and thinks that not-p. Reconsidering the first argument, he finds it still compelling and fluctuates between thinking that p and thinking that not-p. He, in fact, does not know what to believe. However, Bach maintains that we can't describe this case as alternatively believing and disbelieving that p or even as believing p and believing not-p, since no position has been settled on. The person has not yet chosen one

belief, although he is constantly thinking of the alternatives. Bach offers the example of the person who has a fear of flying (1980, p.357). This person believes that flying is safe (at least, as safe as driving a motor car) and yet everytime he has to fly he experiences acute anxiety. Eventhough realizing the irrationality of his thought, he cannot help thinking that flying is dangerous.

Before going on to look at Bach's view of evasion in more detail, I want to note a similar distinction to that of Bach. Instead of distinguishing between believing that p and thinking that p, Cohen makes a distinction between believing that p and accepting that p.

> "Very often we accept what we believe and believe what we accept. But a person who does not fully believe that p can nevertheless accept that p .... Equally, a person can fully believe that p without fully accepting it ... Again, accepting that p is no reason at all for believing that p. But having a belief that p is normally some reason for accepting that p, even though it may well not be the only, or the best, reason or even a sufficient one." (1986, p.92-93)

The first case of accepting but not fully believing that p may refer to a subjective state of conviction. For example, a person has a hunch that Mr X is in fact a swindler although all the presently available evidence and the socially accepted opinion is that Mr X is an honest pillar of the community. In the case of a person fully believing something without accepting it, Cohen uses the example of a juror who believes and is convinced due to a personal acquaintance that his witness is untrustworthy, but he rejects the use of this belief as a premise for certain proofs. In fact, he feels that he should put this belief out of his mind when he is judging the case in court.

I shall now return to develop Bach's account of self-deception as evasion.

After noting the distinction between the belief that and the thought that, Bach concludes that "the self-deceiver, believing that not-p while desiring that p, need not ... try to get himself to believe that p. That is neither his objective nor essential to it. It is enough that he not (sustainedly or repeatedly) thinks what he believes, for what matters is what occurs to him." (1980, p.357). Whereas Martin's view of "weak" self-deception entails an absence of a true belief, Bach's view proposes an absence of a false belief. It is enough for Bach that the self-deceiver who desires a certain state of affairs should avoid thoughts about his belief as to the real state of affairs. According to Bach the self-deceiver does not aim at forming an irrational belief, or does not aim at fooling himself into believing something false. What happens is that his desire for p motivates him to avoid thoughts generated by the belief about the truth, not-p. Here then is a simplified analysis of self-deception:

1. A desires that p
2. A believes that not-p
3. 1 + 2 combine to motivate A to avoid (and he does avoid) the sustained or recurrent thought that not-p.

By simplifying his analysis, I have laid Bach open to perhaps unjustified criticism, but I have done so to highlight two issues, viz. the problem of how A's desire that p and his belief that not-p can combine to motivate him in sustaining his self-deception, and secondly, the different ways in which A avoids the sustained or recurrent thought that not-p.

Turning to the first problem then of how the desire that p and the belief that not-p combine to cause A to avoid the thought that not-p (at least on a sustained or recurrent basis). It might be proposed that the desire and

belief jointly constitute a reason to avoid the thought that not-p but it is still difficult to see how they can constitute a reason for the person to act on. If the self-deceiver reasons as follows: "Although I believe that not-p, since I desire that p, I will avoid the thought that not-p whenever not-p comes to mind," we are returned to the very paradox we wish to avoid. Bach's way out of the dilemma is to state that although the self-deceiver's desire and belief do not constitute a reason on which the agent acts in order to avoid the unwelcome thought that not-p, they do comine to motivate him to avoid that thought. The fact that he cannot bear the thought that not-p is not his reason for avoiding it. The self-deceptive process motivates him to find reasons for rejecting not-p or reasons in favour of accepting p. "The self-deceiver's desire that p and his belief that not-p combine to cause him to accept reasons for avoiding the thought that not-p" (p.366). However, to say that he is motivated to reach the conclusion he does is not to explain just how he does so. Bach's unsatisfactory reply to this very real gap in his theory is, "However, we may reasonably suppose, considering his motivation, that he would have reached this conclusion somehow." (p.366). The self-deceiver is motivated to rationalize his behaviour by finding a reason for it; for example, the malingerer has a desire to avoid responsibility which, in turn, motivates (and eventually causes) him to find a more acceptable reason, like ill health, to rationalize his staying in bed. The self-deceiver is, of course, not forced (e.g. by unconscious defence mechanisms) to avoid the thought that not-p, only motivated to avoid it. He can become undeceived at any time. If, however, he is confronted with what he is doing and denies it, he is doubly motivated: motivated to avoid the thought that not-p (primary deception) as well as being motivated to avoid awareness of what he is doing (secondary deception). On the secondary level of deception his desire for the success of his self-deceptive project and the belief

that he is practising self-deception combine to motivate him to find "secondary" reasons for his behaviour, in order to rationalize his own behaviour. "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 not-p but is unaware (or denies the impact of) this motivation and of his uncharacteristic violation of his own rational standards." (p.368). The puzzle, however, remains of how one can be caused to accept reasons one would normally reject. Although Bach does not attempt to provide a detailed solution to the puzzle, he maps his view onto that of Pears by stating that the self-deceiver relies on "the boyancy of his wish" (Pears, 1984, p.110) to sustain and defend the irrationality of his actions. (Pears' view will be examined in Chapter 5).

I now want to turn to the second issue raised by Bach's analysis, viz. the different ways in which A avoids the sustained or recurrent thought that not-p.

The first tactic is that of simple evasion. This is perhaps the simplest of the psychological devices that Bach mentions in that it involves no more than evading the thought that not-p by thinking of a single reason against not-p. The reason is not conceived as proving or disposing the self-deceiver to make a temporary judgment, it is simply a reason for getting his mind off not-p. The reason he has settled on need not be a convincing one - all it needs to do for it to be accepted as a reason is for the self-deceiver to stop thinking any further on the issue. If he does think further on the issue and the alleged reason for it, he may have to resort to rationalization —— a tactic which will be examined later on in this chapter. For Bach, simple evasion is merely "turning one's

attention away from some touchy subject", (p.360) so that one evades having to admit to oneself what one believes. Bach compares this self-deceptive process to procrastination. "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". (p.361). For example, the girl's desire for her lover's faithfulness and her belief that he is unfaithful motivate her to accept a reason for evading the whole question of her lover's infidelity. She may accept her busy work schedule as a reason for avoiding the whole issue and if her friend should invite her for lunch at the café her lover frequents, the girl may plead too much work and so avoids being perhaps confronted with damaging evidence. The self-deceiver's reason may seem contrived, but more efficient than the pseudo-reason itself is the thought that there is some such reason —— identifying it is unnecessary. Just the thought that there is such a reason serves itself as a reason. "Most efficient is the thought that not-p is not worth thinking about. What better reason not to think about it?" (p.361).

The second psychological device Bach mentions is that of "jamming", which is also a kind of evasion of the truth and of the belief that not-p. This consists of cluttering one's mind with considerations in favour of p. Whenever the issue of not-p comes up, the self-deceiver focuses his attention on p, as a result of his desire that p, and imagines what it would be like if p were true or he vividly imagines desirable consequences of p. He may run through evidence favouring p and perhaps even go so far as to provide himself with instant "evidence" for p. This can be done either by acting as if p were the case (the self-deceiver employs the strategy that operates through output - see Chapter 3) or by convincing

others that p and then taking their word for it, or both. By such means as these the self-deceiver, when confronted with not-p which threatens to bring to his attention his belief that not-p, clutters up his mind with the thought that p, and so "jams" or prevents his belief that not-p from coming into his focus of attention. Bach offers the example of a young man who resents having to care for his aged mother. The fact that he experiences feelings of extreme intolerance and resentment is evidence enough for the belief that he hates her (the belief that not-p). However, he does not want to hate her (the desire that p) and so, whenever the subject of his resentment crops up (his mother brings it up often enough), he clutters up his mind with nice thoughts about his mother, brings her roses, pays her compliments (especially when others are present), etc., so that there is no room in his attention for the thought that he hates her.

Both in the tactics of evasion and jamming, the self-deceiver has the belief that not-p as well as the desire that p. These two aspects do not necessarily give rise to a contradiction and Bach offers two ways in which the paradox is avoided. Firstly, the self-deceiver need not come to believe that p (the desire that p which motivates him to avoid not-p is enough), thus we are not confronted with two conflicting beliefs, the basis for the epistemological paradox. Secondly, the self-deceiver is not aware of his belief that not-p when he doesn't think about this unwelcome belief, and so the self-deceptive tactic is to avoid the thought that not-p which will bring to his attention the belief that not-p. If the self-deception is successful, the belief that not-p does not come to mind and there is, therefore, no paradox of how the self-deceiver can believe p and believe not-p at the same time. Bach's analysis is not without its attractions, but it does not cover all cases of self-deception, e.g. when a person is

confronted with such conclusive evidence that he cannot sanely evade or ignore the issue. Furthermore, the analysis does not offer a comprehensive account of the mechanics of avoiding the thought that not-p. Is it conscious thought to avoid not-p? If so, then both the thought that not-p and the thought to avoid not-p must be in the focus of attention: a seeming return to the paradox. Is it an unconscious thought to avoid not-p? If so, then how was this thought suppressed into the unconscious and how does it manipulate the consciousness of the person? That the person avoids thinking about the unpleasant truth in a self-deceptive project is obvious enough, but just how it happens is not so obvious, that is if Bach's view of self-deception is accepted.

I now want to turn to another interpretation of self-deception which hopes to avoid the paradoxes, viz. failure to focus which also encompasses one-sided evidence gathering and biased thinking. The tactic of failure to focus in self-deception is very similar to Bach's evasion in that both employ the strategy of filtering what gets into the mind. Here the self-deceiver's wish that p may lead him to fail to focus his attention on evidence which he has for disbelieving that p or for believing that not-p. Whenever the self-deceiver becomes aware of contrary evidence, his attention is diverted to other things. This diverting of attention may be either intentional or unintentional. For example, whenever the girl's attention is attracted to evidence of her lover's unfaithfulness, she may tell herself that it is a waste of time to consider this evidence, since her lover is just not that sort of person to do such a thing. This, of course, is very similar to Bach's account of simple evasion —— an intentional diverting of attention by thinking of a reason against the suspected unwelcome belief. The desire for his faithfulness may be so strong that whenever she is confronted with evidence for the belief that not-p, she automatically

shifts her attention to other issues. The small difference of this account to that of Bach's is that whereas Bach's self-deceiver avoids the thought that not-p by thinking of one reason against it or by cluttering up his mind, this failure to focus means that the self-deceiver shifts his attention to other things. By always shifting the attention to other things, this process eventually becomes automatic. The desire to believe that p is so powerful that it influences the workings of her mind by either shifting thoughts away from the dangerous evidence for the belief that not-p or by preventing her mental processes from "thinking through" the implications of the evidence which seems to support the belief that not-p. According to Mele (1983, p.372), she does not intentionally aim at failing to focus on the evidence, but the desire for pleasure in the belief that her lover is faithful "automatically" shifts her thoughts from the unpleasant suspicion that he may be unfaithful. An obvious objection is immediately evident: just how does this "automatic shifting" work? To this central question Mele offers no answer, but he tries to circumvent the epistemological paradox by maintaining that failure to focus on the evidence for not-p does not in any way imply that the self-deceiver already holds the belief that not-p. For example, the hapless girl may have had a disastrous love affair in the past and has vowed that she will never again try to find out whether her future lovers are faithful or not. She prefers not to know what he does when he is out late —— ignorance is more pleasurable than knowing (whatever the outcome). This does, of course, evade the paradox posed by Sartre that one must know the truth very exactly in order to avoid it. The girl has no exact knowledge of the truth but avoids that entire area which may prove to be distressing. For Bach, the belief that not-p only comes to attention when the self-deceiver has the thought that not-p, whereas in this account the belief that not-p is absent altogether. In this way, failure to focus circumvents

the epistemological paradox of p and not-p, in that the belief that not-p need not even arise. However, this view —— that the failure to focus on evidence for not-p in no way implies that the self-deceiver holds the belief that not-p —— may solve the paradox of holding two conflicting beliefs, but the account at the same time robs self-deception of its distinctive characteristic —— that of mental conflict. It is difficult to see how the above interpretation of self-deception differs from that of wishful thinking. Furthermore, this interpretation, like Bach's, cannot account for the case in which the evidence for not-p is so strong that it cannot be ignored by a sane, rational person.

Coupled to failure to focus is the psychological device of one-sided evidence gathering, employing the strategy which operates through input and which, therefore, results in biased thinking. However, one-sided evidence entails that the practitioner holds the belief that not-p and it is the very knowledge of the truth of this belief that steers his manipulation and exploitation of the data. There are eight different ways in which the biased practitioner can deal with evidence:

1. to look for or highlight evidence that supports the favoured belief that p
2. to look for or highlight evidence that refutes the unwelcome belief that not-p
3. to avoid or ignore evidence that refutes the favoured belief that p
4. to avoid or ignore evidence that supports the unwelcome belief that not-p
5. to interpret certain evidence as not counting against the belief that p when in actual fact it does.

6. to interpret certain evidence as supporting the belief that p when in actual fact it does not.

7. to interpret certain evidence as counting against the belief that not-p when it actual fact it does not.

8. to interpret certain evidence as not supporting the belief that not-p when in actual fact it does.

The slight difference in approach to evidence in 1 - 4 and 5 - 8 is that in the first four ways the self-deceiver looks for (or ignores) certain evidence and once he has found it (or successfully avoided it) he uses it (or the absence of it) as is. In the latter four ways the self-deceiver is confronted with certain evidence which he then changes and misinterprets. For example, a theologian holds a certain political view and wants to believe that this view is "ordained by God". He consequently scours the Biblical texts for evidence, overlooking obvious evidence to the contrary, and finally succeeds in finding obscure ambiguous (albeit not to him) evidence to "support" his belief that p. Or he may use some well-known passage in the Bible and give it a "novel" interpretation, an interpretation that will, of course, lend support to the belief that p. Whichever way he deals with the evidence (ignoring/highlighting or distorting evidence), he is guilty on at least two counts of irrationality: firstly, he is irrationally "hypersensitive" to certain obscure details, or irrationally "blind" to counter evidence, or irrationally misinterpreting evidence (primary deception); and secondly, as a trained historian, he should be well aware of the illegitimacy of this "method" of research, i.e. he should be aware of his own irrationality (secondary deception). The biasing agent is, of course, his wish for the belief that p. In order to promote the favoured belief that p (or to refute the unwelcome belief that not-p) the wish directs the self-deceiver's interpretation of evidence in such a way that only selected

evidence which is wanted by the wish is allowed into the mind. Once the one-sided evidence has been registered in the mind, biased thinking takes over.

Biased thinking is often taken as being the same as one-sided evidence gathering, but there is a difference. The latter device operates on input, whereas biased thinking works on what is already in the mind. The two strategies as well as the two psychological devices are mutually re-inforcing. The agent concentrates on selecting only that evidence which he wants, which in turn gives rise to biased thinking within the mind. Once the agent has the biased belief he will then direct his attention only to that evidence which supports it. Whereas the one-sided evidence gatherer approaches data mainly in ways 1 - 4, the resultant biased thinker interprets the data mainly in ways 5 - 8. (As I have mentioned, the two devices are mutually reinforcing and there is, therefore, no specific separation in the ways 1 - 8. The different approaches are often used interchangeably).

Bach points out that self-deception 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." (Bach, 1981, p.352). The distinction then between plain biased thinking and self-deception seems to point to the rôle the agent plays in bringing about his irrational belief. In plain biased thinking the agent plays a fairly passive role ——his intellectual processes are inflexible, or he is intellectually incompetent to appreciate the importance of counter-evidence. However, in self-deception, the agent seems to play a more active role——his desire motivates him to actively misinterpret

the evidence to suit his own desired belief, or to ignore damaging evidence or to look for supportive evidence. The self-deceiver, under the influence of a wish, makes a deliberate effort to support the irrational desire-goaded belief, through the distortion of his rational processes—— an action which, in a rational person, gives rise to the psychological paradox. The theologian is not merely politically biased——he actively hunts for evidence to support his desired belief and devises elaborate "novel" interpretations of the data. Of course, the more active a rôle the agent plays in avoiding or distorting the evidence, the more radical the degree of self-deception. To be totally "blind" to certain damaging evidence, in such a way that does not seem possible, is a far more severe form of self-deception than the person who pays scant attention to evidence of which he should really take note. Elster cites a case in which millions of people were, to an extent, self-deceived in that they deliberately refused to confront threatening evidence. During World War II millions of Germans deliberately overlooked the extermination of the Jews:

> "(T)hey must have observed that their Jewish acquaintances disappeared, and they must have known that this had some gruesome explanation, but as long as they managed to remain ignorant of the details they could say to themselves that they were genuinely unaware of what went on."
> (Elster, 1979, p.178)

This is an example of one-sided evidence gathering (case 4) which does not, however, lead to biased thinking since it is calculated to keep them in a state of ignorant bliss. This is not a paradoxical case of self-deception, since they did not have certainty of the facts. Instead of the belief that not-p, they may have had a suspicion that not-p. If they did have knowledge, it was knowledge that such facts did exist, but they had no knowledge of the contents of those facts. If we are to accuse the Germans of irrationality, we can do so in Davidsonian terms, in that their irrationality lies in their

refusal to collect all available evidence, but not in their holding two incompatible beliefs. (Those Germans who knew the contents of the facts and yet deceived themselves about the holocaust are, of course, guilty of a more radical type of self-deception, a type that would need to employ a more radical psychological device if the self-deception is to be preserved. These more radical types of psychological devices will be examined as the discussion moves along the sliding scale of self-deception).

Another explanation of how it is possible for a person to have two conflicting beliefs within the same belief system is that of mental distance. It aims at providing a solution to the following problem: the agent is motivated by his belief that not-p, is therefore aware of not-p, and yet this awareness leads him to persuade himself to believe that p. The case of having incompatible beliefs can, to a certain extent, be compared to that of having incompatible desires. Of course, we all experience conflicting desires in our life ("I want that dress", "I don't want to pay so much for it") and we may even experience incompatible desires ("I want to drink" and "I don't want to drink" ——I am thirsty but the only liquid available is cold, unsweetened coffee). The case of incompatible desires need not constitute a paradox (see footnote (1)), but the case of incompatible beliefs is somewhat more problematic since beliefs aim at the truth. If there are inconsistent beliefs within the same belief system it means that not all the beliefs can be true. A rational person, aware of the inconsistency, will change his beliefs to fit the facts. However, the self-deceiver does not reach this stage. He keeps the two inconsistent beliefs in his belief system but, as a rationally competent person, he cannot allow himself to become aware of the inconsistency, something that would be all too obvious if he should try to conjoin consciously the two incompatible beliefs (a seemingly impossible task) or even if he held both

beliefs in the same focus of attention. How he can circumvent becoming aware of the incompatibility within his belief system is to keep the two incompatible beliefs apart. The agent does not, of course, do this consciously (that would defeat the self-deceptive project before it started), but the agent's wish to believe that p ensures that the belief that p and the belief that not-p are kept apart through mental distance. For example, an educator believes that corporal punishment is bad and during lectures he argues most sincerely and convincingly against corporal punishment. Yet, when he is at home, he frequently gives his naughty young son sound hidings. This example does not show how the incompatibility disappears, but shows that people can, quite sincerely, hold two incompatible beliefs. The form of the epistemological paradox which is applicable here is: aBp+aB-p, with mental distance separating the two.(5) What allows the rational educator to keep the beliefs apart can be an inability or disinclination to think things through (this then is rather a case of irrational belief-formation than of self-deception) or his desire for p which motivates him to keep the belief that p and the belief that not-p apart (a case of self-deception). If, however, after someone has pointed out this inconsistency, he still clings to both beliefs, we are faced with a more problematic case of self-deception in which he would not only have to conceal from himself the belief that not-p when he is at home, and to conceal from himself the belief that p when he is lecturing (i.e. primary deception) but he will also have to conceal from himself the belief that he is holding two incompatible beliefs (i.e. secondary deception). Using the strategy which operates on material already in the mind, he may resort to self-deceptive tactics like rationalization in which a pseudo-reason is fabricated in order to rationalize his action of beating his son. This notion of mental distance does not solve the epistemological paradox —— it merely shows that the self-deceiver can hold two incompatible beliefs in the same

belief system, but the above interpretation does not show how mental distance functions, i.e. the mechanics of mental distance. Problematic questions arise: how does the wish manage to keep the two beliefs apart? and is there, on a second-order level, another belief that one must hold the two beliefs apart? However, I shall not dismiss the notion of mental distance altogether. In Chapters 5 and 6 I shall again refer to mental distance, but in the employ of a more radical type of self-deception (cases 3 and 4) and shall look at the mechanics which keep the two incompatible beliefs apart.

Haight argues that whether A can simultaneously believe that p and believe that not-p depends on:

> "..... whether A can do two different things at the same time, each of which counts independently as actual belief. For example, if merely acting as if were enough to count and so was sincere and intelligent assent even when unspoken, A would need only to act as if p with a sincere mental reservation that not-p and the thing would be done." (1980, p.18) (4)

Here self-deception moves into the realm of rôle-playing which is an all too familiar form of "weak" self-deception, like the unsavoury character who acts as if he is (and he sincerely believes himself to be) the hero of every female's fantasies. The fact that he is spurned by every woman he meets does not seem to deter him in the slightest nor to change his irrational belief. Self-deceptive rôle-playing does not refer to doing a job (like an actress who "lives" herself into the character she is portraying) or to amuse oneself or others. Self-deceptive rôle-playing employs the strategy that operates through output, "the backward connection", and has as its aim to inculcate a certain belief in the player which will override other beliefs that we think he must have and should acknowledge. At the beginning of Chapter 1 I listed various "typical" cases of self-deception,

including the one of the physicist who refuses to acknowledge that his research in nuclear arms contributes to warfare in general. He may justify his belief by evaluating himself purely in the rôle of a scientist — i.e. a discoverer of knowledge. He may identify himself as nothing but a scientist, which then absolves him of the political or social repurcussions of his research. Or, at least, he identifies himself solely in terms of his scientific rôle as far as the ethics of his work are concerned, i.e. the advancement of knowledge. By acting the rôle of the scientist purely concerned with intellectual discovery, he may come to get the belief that his responsibility as a scientist is solely that of intellectual discovery, regardless of the consequences of that discovery.

Of course, one of the best known examples of self-deceptive rôle-playing is that of Sartre's waiter (1958, part I, ch.II, sec.2). The man plays the socially expected rôle of a waiter while he is attending to the diners. The man is dictated by the part he and society see him in. Since he has not freely chosen to play this rôle as he does, his behaviour carries a stamp of artificiality — he is a little too quick, too eager, too solicitous. This kind of rôle-playing denotes an aspect of bad faith, or self-deception, since the waiter believes that he is determined by the rôle socially assigned to a waiter. In fact, according to Sartre, he could at any moment choose to behave otherwise, but he manages to obscure himself from the truth, namely that at any moment he could stop behaving as he does. He evades the responsibility of "being-in-itself" — he denies that he can be other than he pretends to be. Instead, he sees himself as caught in the rôle of a waiter in the mode of "being-what-he-is-not." His desire to escape his responsibility motivates him to form and accept the comforting belief that he is nothing but a waiter, a rôle in which the patterns of behaviour are socially determined. And by behaving like a waiter should,

he provides support for the belief that he is nothing but a waiter, a self-deceptive belief which absolves him from taking responsibility for his own actions. Since this is a rôle forced on him, he also absolves himself of the responsibility for choice and, he believes, of the capacity for choice. With no capacity for choice, he is no longer troubled by the irrationality of his actions, since it is not something he has brought about; it was forced on him. The waiter, therefore, escapes from the anguish of the responsibility that his acts are totally hiw own, via bad faith, by inventing some kind of psychological determinism which forces him to act as he does. Phillips makes (1981, p.30ff) an interesting observation. According to the Sartrean view then, an effective waiter is necessarily a victim of bad faith in that the job prescription of waiting on tables necessarily involves duties which make it impossible for the waiter to be himself — e.g. to be courteous when, in fact, he is annoyed; he has to smile when, in fact, he dislikes the customer, etc. There can, therefore, according to Sartre's argument, be no authentic choice to be a waiter.

The tactics of both mental distance (keeping two conflicting beliefs apart) and rôle-playing can be employed at the same time by the self-deceiver. These two tactics are not always separate, but are often used by the self-deceiver to re-inforce his irrational belief and behaviour. The combination of the two tactics re-inforce the irrationality as follows: because he believes that p he acts in a certain way, and because he acts in a certain way it must mean that he believes that p. The rôle-player can draw upon two different belief systems, each rôle calling upon and generating a specific belief system. Of course, the bulk of the belief systems overlap, e.g. he believes that he lives in Cape Town regardless of what rôle he is playing (whether that of faithful husband or that of carefree bachelor),

but through rôle-playing he can keep the irrational action coupled with the irrational belief apart from the other rôle coupled with the conflicting belief. The same criticism of mental distance applies to rôle-playing in that there is no explanation of how the schism between two rôles is maintained, or of the mechanics of a possible dominance by the prefered rôle. (A complete take-over by the irrational rôle, however, seems to steer us into the designated area of multiple-personality, such as the well-known cases of Eve and Sybil). The rôle-player, however, has the added advantage over the self-deceiver who employs only mental distance as a self-deceptive tactic: through behaving as if the irrational belief were true, the rôle player re-inforces his belief and supplies himself with concrete evidence (i.e. his actions) that supports the favoured belief, draping it in an artificial cloak of rationality.

It is interesting to note the idea of rôle-playing in its extreme form, i.e. split personality. Haight cites instances (1980, p.133ff) where one of the "selves" of the split personality had privileged access to the other "self" but not vice versa. In such a case it is, of course, possible to conceive literal self-deception in its most problematic form, i.e. one self deceiving the other by manipulating its consciousness. Here we encounter two "selves" housed in one body. Since the two "selves" seem to have two different personalities and belief systems, the one not having access to the other, it is easy to see how self A_1 can deceive self A_2. A (comprised of A_1+A_2), therefore, can believe that p and can believe that not-p.(6) Because "multiple personalities" are far removed from typical cases of self-deception, I am not going to deal with the obvious objections that can be raised about split personalities (e.g. can the two "selves" really count as "a" self that deceives "itself"? and is the belief that p in the consciousness of A_1 or in the unconscious of A_2 when A_2 is

"operating"? etc.). We may learn something from these psychological exceptions about the nature of self-deception, but what is of real interest is the fact that rational, sane beings are able to deceive themselves.

I shall now return to the theory of cognitive dissonance that I first mentioned in the previous chapter. Although cognitive dissonance is in itself not a psychological device for self-deception, it is of importance for it is a motive for self-deception, an especially strong motive for the use of selective exposure to information (the same as one-sided evidence gathering) and for rationalization. Before looking at the important device of rationalization, I shall therefore return to the findings of the Festinger school of cognitive dissonance and will show how rationalization has been employed by the subjects in the experiment in order to reduce cognitive dissonance.

> " (Cognitive dissonance is) the jarring relation that holds between beliefs, or between beliefs and actions, when they belong to an irrational sequence ... According to dissonance theory, if a person says something he feels is untrue, he experiences dissonance: the cognition 'I said $\underline{x}$' is dissonant with the cognition 'I believe not-$\underline{x}$'. In order to reduce dissonance, he might attempt to convince himself that what he said was not so very untrue." (Pears, 1982, p.280)

If we return to the experiment in which the students had to lie (see Chapter 3), it seems that the students, in order to reduce dissonance, look for justification of their saying $\underline{x}$. The results of the experiment show that the greater the justification (e.g. a great deal of money was paid for telling the lie) for saying the opposite, $\underline{x}$, to what they believe, not-$\underline{x}$, the less dissonance between the act of lying and the belief that the lie is a serious one. Interpreting the cognitive dissonance theory in this

way seems to steer my discussion in the direction of akrasia, acting against one's own better judgment (the jarring relation between final judgment and action), whereas I am really concerned with self-deception, the jarring relation that holds between beliefs. However, the above experiment is applicable to self-deception as well. First of all, the students experienced dissonance due to the conflict between the belief that lying is wrong and the awareness of having told a lie. To decrease dissonance they may proceed in two different ways: first of all, they could "suppress" the awareness of their akratic act. This, however, is a drastic tactic, one whose success is not guaranteed——the awareness may surface at any time, doubling the dissonance experienced: that is, the dissonance created by the "re-surfaced" awareness of the akratic act as well as by the awareness that he was trying to suppress the awareness of having told the lie. This form of self-deception which employs the psychological device of suppressing an unwelcome belief is certainly not impossible, but it constitutes a more problematic case of self-deception which will be discussed later. The other approach to decreasing dissonance is to tackle the other part of the conflict, viz. the belief that lying is wrong.

The student's desire to believe that he has not done wrong (or that what he has done is not so very wrong) motivates him to look for reasons that will justify him to change the belief that lying is wrong. To change the belief does not necessarily mean to negate it. Change can imply a modification of the belief. The great deal of money that was paid is the justification he concentrates on in order to change the belief that lying is wrong to the belief that some lying is not so wrong. He may reason that the money he received will be put to good use (e.g. a donation to a morally worthy cause) and since the consequences of the lie are good, the lie

itself cannot be all bad. The initial belief "lying is wrong" which brought about dissonance has been camouflaged into the belief "some lying is not so wrong", thereby reducing the dissonance brought about by the awareness of his akratic act. It may be argued that what happens here is merely that a large temptation excuses a wrongful action. However, there is a difference. By merely appealing to the large temptation of the money, the student may excuse his behaviour, but he doesn't try to change his belief that he did wrong. He exonerates himself from responsibility by casting the large temptation in the rôle of an outside force — one that overwhelmed his better judgment. The self-deceiver, on the other hand, tries to reduce his cognitive dissonance by changing his belief that he did wrong. The reduction in dissonance can be set out as follows:

1. He holds the belief "lying is wrong"
2. He has the awareness "I lied"
3. As a rational person he, therefore, concludes, "I did wrong." This conclusion is in direct conflict with
4. the desire to believe, "I am a good person" and he, therefore, experiences great dissonance.

In order to reduce the dissonance he embarks on a self-deceptive project:

5. by exposing himself to selective information sources, he deduces from the biased evidence the belief "lies with good consequences are not wrong"
6. He re-evaluates "I lied" to "I told a lie that will have good consequences"
7. and reaches the conclusion, "I did not do wrong", an opposite belief to the one in (3) and one which is not in conflict with
8. the desire to believe "I am a good person", and so the dissonance is greatly reduced.

According to Pears, the really uncomfortable dissonance is that created by the akratic act, i.e. the act of lying and the belief that the lie is a serious one. "The strategy of self-deception in such a case is to reduce the really uncomfortable dissonance by creating a less uncomfortable dissonance, which is, in any case, concealed from the subject because self-deception is usually a self-covering operation." (Pears, 1982a, p.280). For Pears, then, the most important device is that of self-deception in the employ of akrasia, that is to reduce dissonance. This first part is, therefore, primary deception. But a problem is immediately apparent: as a rational being he is aware of the jarring relation that holds between his belief and action and tries to mask this "irrational sequence" with a rational facade. To bring this about he calls on the services of self-deception, especially those of rationalization. But the self-deceptive project can have the opposite effect on him: instead of decreasing dissonance between the action and belief, it can increase dissonance between his beliefs. As a rational person, the self-deceiver must be aware of the jarring relation that holds between his beliefs—— between his information about the circumstances of the lie and his minimization of its seriousness. In other words, the dissonance experienced as a result of his awareness of his self-deception calls for a secondary self-deceptive project to reduce dissonance caused by awareness of the first project. For Pears, however, this awareness constitutes the "less uncomfortable dissonance," a dissonance that is dissipated by secondary deception, "a self-covering operation."

Now that cognitive dissonance, as a motive for rationalization, has been discussed, it is necessary to look at how the self-deceiver goes about looking for justification of his irrational belief or his akratic action. Rationalization can take various forms: the self-deceiver may "fool himself"

about his real reasons for an act by evading the real reasons, or he may try to justify a belief or act by finding "good reasons" for it, by fabricating pseudo-reasons. In the latter case, he, therefore, looks for "rational" justification by constructing a pseudo-reason that supports his favoured belief. He can look for these pseudo-reasons either by the aforementioned process of selective exposure to information, or he may fabricate a pseudo-reason that is totally unfounded by any evidence. Either way, the method of obtaining the pseudo-reasons is irrational. However, rationalization can extend to the general case of a person's explaining away what he would normally regard as adequate evidence for a certain proposition, but, as Bach points out (1980, p.358) there is nothing intrinsically irrational about explaining away evidence against what we already believe. Bach notes that "this is part and parcel of good scientific method——theories should not be discarded without a fight——and of everyday thinking as well."(p.358). We often look for errors in "recalcitrant experiences" rather than adjust our beliefs immediately and without question, but the boundary which separates this rational practice from the irrational practice of the same method is vague. Perhaps the work of intrinsic irrationality can be found in the practice of explaining away evidence against a belief purely because the evidence weighs against what one desires, especially if in the absence of that desire one would adjust one's belief to the evidence. "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." (p.359). The rationalizer's "reasoning" aims at convincing him that he is justified in holding the belief that p. As a rational person, he knows that one does something because one has a reason for doing it, and by evading the real reason, he is able, through rationalization, to provide himself with another more acceptable reason for his akratic action or irrational belief.

Typical rationalizations include "sweet lemons", e.g. a girl, whose hero left her for another, marries "the next best" and exaggerates his merits to show others and to convicne herself that she really got what she wanted. Another typical form of rationalization is "sour grapes" in which the girl may play down her attraction to the hero to such an extent that she persuades herself that she never wanted to marry him in the first place and that she is actually glad to be rid of him. The rationalizer may also blame circumstances (e.g. the bad workman who blames his tools for his own shoddy work) or other people (e.g. the student who fails his driving test blames the examiner for his failure by claiming that the examiner had a personal grudge against him).

Pears notes a phenomenon that arises from the cognitive dissonance theory, viz. that of "spreading". (Pears, 1984, p.57). When someone has come to a conclusion after long deliberation between two "close run" beliefs, he is likely to feel uncomfortable about the precariously held final choice. In order to rid himself of this uncomfortable feeling, he may exaggerate the considerations that favoured it and minimize the considerations that counted against it. Although this form of rationalization is very similar to that of "sour grapes" and "sweet lemons" in which the person has to make the most of something that was forced on him, "spreading" differs slightly in that it is aimed at reassuring the person that he was not mistaken in the choice he made voluntarily.

The cognitive dissonance I am interested in, especially as a motive for rationalization, is the tension, amounting to inconsistency, amongst a person's beliefs and attitudes. Even though all the above forms of rationalization seem to involve fooling oneself about one's real reasons or fabricating pseudo-reasons for the akratic act, rationalization aims at making

the belief about the reason for the action more acceptable to the self-deceiver. The self-deceiver does not fool herself about the nature of the actual action— she does not deny that she no longer has her "hero"— but she fools herself about the belief about the reason for the termination of the relationship. Pears notes that another way of reducing the dissonance is not only to rationalize the belief about the reason for the action, but also to rationalize the belief about the actual nature of the action. The self-deceiver can change his belief after the action and then evaluate his action in terms of the new belief, leading him to change his mind about the wrongness of the action. (The step-by-step analysis of how he accomplishes this was noted earlier on.) In the example of the students, a more drastic form of rationalizing the belief about the nature of the action, a form of "white washing" the lie, was the claim by the students that they had actually held the changed belief before they told the lie. The students' wish to be rational is the motivating factor for this more drastic form of "white washing". It is this very wish to be rational that directly produced the belief that they had been rational by acting in accordance with an already held belief, and not a belief that was formed after the act.

I have devoted quite a large section to cognitive dissonance and the self-deceptive tactic of rationalization in order to decrease dissonance, but how exactly do these issues confront the paradoxes of self-deception? Other theories of self-deception explain away the epistemological paradox through either an absence of the true belief that not-p (e.g. evasion, jamming, failure to focus) or a lack of certainty about the truth of the belief that not-p or falsity of the belief that p (brought about by a refusal to collect all the evidence) or by keeping the belief that not-p and the belief that p apart in the mind so that the self-deceiver need not

become aware of the contradiction within his belief system (through mental distance and rôle-playing). Rationalization, however, aims at changing the belief that not-p so that it is more in line with the belief that p. The closer the belief that not-p (e.g. "lying is wrong") can be brought to the belief that p, the less apparent and uncomfortable the contradiction (e.g. "some lying is not so wrong"). Through selective exposure, *he* finds justification for changing the belief that not-p. Again, I want to stress that rationalization is not usually used as an isolated tactic of self-deception. More often than not, in real-life situations, rationalization is backed up by other self-deceptive tactics like evasion, mental distance, failure to focus, etc.

I want to illustrate this point with an example from Gide's Symphonie Pastorale. Palmer (1979b, p.88ff) also refers to the pastor who deceives himself into believing the opposite of what he knows to be the truth. The pastor is in love with the blind girl Gertrude who has been placed in his care. After returning from a concert with her, he is reproached by his wife that he does things for Gertrude which he would never have done for his own children. He deceives himself by thinking that his wife's reproach was unfair because she knew perfectly well that the other children were occupied in some other way, and Amélie, his wife, has no interest in music in any case. The pastor deceives himself as follows: his wife's alleged indifference to music and his children's prior engagements are actual facts which he construes as the reasons for failing to invite them when, in fact, the real reason is that he wished to be alone with Gertrude. His failure to invite his family along (which points to his wanting Gertrude to himself and his feelings of love for her) is rationalized by claiming that because the others were busy and uninterested in music he did not invite them and there was, therefore, nothing else for him to do but go alone with

Gertrude. The pastor, thus, provides himself with another more acceptable reason——the rationalized reason——for his action. Rationalization in this way reduces cognitive dissonance in that the pastor persuades himself that his action of failing to invite the others (or rather his belief about the action) has a justified cause. Rationalization is normally used by the self-deceiver after the action has been done or after the evidence has been confronted, but the self-deceiver can also "backdate" the newly rationalized belief about the reason for the action to before the action took place so that this "justified" belief would make the action appear to be rational. The pastor evades questions that might make him aware of the real reason for failing to invite them (i.e. the desire to be alone with his beloved), by filling his mind with other considerations. He admits that after the concert he forgot to buy the cotton reels Amélie had requested but, he tells himself, he is more vexed with himself for this small oversight than she could ever be. Had not his conduct been beyond reproach? By cluttering his mind with these sanctimonious observations, he deceives himself by ignoring the painful questions which, had he been honest with himself, he would have put to himself. He ignores questions about the nature of his relationship with Gertrude or whether it was right to make the visit in the first place. The parson, by concentrating on "safe" issues, blinds himself to certain things that are obvious to the reader, as well as blinding himself to certain things that should be obvious to him. Martin points out that most people have some concern for goodness but that often other motives or desires prompt actions inconsistent with that concern. Rationalization "provides an easy way to engage in wrong doing while making our sins 'sit a little easy' on our minds." (Martin, 1986, p.35)

I now want to turn to Fingarette's account of self-deception which, he believes, escapes the paradox. He proposes a different approach to the

analysis of self-deception, "one which does not centre on the co-existence of inconsistent beliefs, and indeed does not centre on the understanding of self-deception in terms of belief at all." (1982, p.212). Fingarette is not the first to approach human judgment in terms of volition. Hume, as cited by Cohen (1986, p.93), states that "belief is more properly an act of the sensitive, than of the cogitative part of our nature." Fingarette argues, in outline, as follows: many of the foregoing accounts are misleading because they are couched in the wrong terms. They are couched in what he calls "cognition-perception" terms. The epistemological paradox arises because self-deception is defined as "believing what one knows to be false", or "consciously holding two conflicting beliefs at the same time but deliberately not seeing one of them", or "a conflict state in which one partially satisfies the criteria for both belief and disbelief", or "buried knowledge." Instead of asking questions about a self-deceiver like "If, in his heart the self-deceiver 'knows' does he really know?", i.e. questions in cognition terms, Fingarette suggests that we should ask "How is he engaged in the world?" and "Does he express this engagement explicitly?", i.e. questions in volition-action terms. Fingarette suggests that, while the concept "consciousness" can be retained, it should be split from the cognition-perception family of concepts, and be treated as part of the "volition-action" family. Consciousness for Fingarette should be thought of as an active mental skill rather than a passive mental mirror. Consciousness, then, is the exercise of the skill of "spelling out" some feature of the world as we are engaged in it. Fingarette concludes that "it is when we judge that there is purposeful discrepancy between the way the individual really is engaged in the world and the story he tells himself that we have the complex but common form of self-deception in which we are interested." (1969, p.63). In other words, the self-deceiver is engaged in the world in some way, and yet he refuses to identify himself as

one who is so engaged; he refuses to acknowledge the engagement as his. For Fingarette, therefore, self-deception turns upon the personal identity one has accepted of oneself, rather than the beliefs one has.

Before going on to look at Fingarette's account of personal identity, I want to note Hanson's explanation of self-deception which takes as its starting point Mead's idea than an individual perspective is tenable only if it functions within a set of perspectives found in the conduct of the community. For Hanson, attributions of self-deception are connected with breakdowns or difficulties in the organization of perspectives, especially tensions between an individual's perspective and that of the community. For Hanson, therfore, as for Fingarette, self-deception centres on personal perspectives rather than on intellectual beliefs. "If the self-deceiver is described as one who holds inconsistent or incompatible beliefs, the special puzzle of self-deception is lost; or, rather, self-deception is not captured in the description, as its puzzle has not been found: Who but a Descartes would think possible the task of sorting out all of his or her beliefs?" (Hanson, 1986, p.109). Very few people are skilled enough logicians to recognize inconsistencies within their belief systems, or to know the full extent of the implications of their beliefs. This observation seems to return us to Davidson's point which I noted earlier on: i.e. the distinction between internal and external irrationality. Hanson's self-deceiver may be externally irrational, but it is only when the agent is aware not only of the beliefs but also of the fact that they, together, are inconsistent, that we have the special and problematic case of inner irrationality.

Fingarette states that a personal identity is established in some respects through the individual's avowing or accepting certain engagements as his.

The self is a synthesis, an achievement by the individual, something he has "made". Fingarette substantiates the creation of the self as a synthesis by tracing specific forms of engagements in the world by the developing child. From rudimentary engagements such as opening a door, etc., he later, as the engagements become more complex, learns about emotions and morals. The engagements blend into one another as a "coherent self" emerges.

> "Certain forms of engagement or even some particular ones are taken up into the ever-forming, ever-growing personal self, and they are modified as they become more and more an integral part of this 'synthesis'. .... To take something into the self is an 'act' which our notion of personal identity presupposes .... If there were no such thing as a person's acknowledging some identity as his and certain engagements as his, and disavowing other identities and engagements, there would be neither persons nor personal identity." (Fingarette, 1969, p.86-87)

According to Fingarette, self-deception then occurs when the person's engagement is unacceptable to himself or when the engagement conflicts with his "self-synthesis". If he were to avow it as <u>his</u> engagement, it would lead to intensely disruptive and distressing consequences. Since the engagement is utterly incompatible with the currently achieved synthesis of self the agent has, the engagement is an autonomous project; one that seems to exist in a certain isolation from the self. The engagement must be divorced from the complex unity of the personal self, otherwise the person's usual reasonableness and values would have a tempering effect on the engagement. Fingarette's explanation of self-deception, therefore, is that it is a refusal to avow one's engagement in the world, due to a painful conflict between one's moral image of oneself and the actual pattern of one's conduct. To resolve the conflict, while continuing the course of action in which he is immorally (according to the self-synthesis) engaged, the self-deceiver denies that it is <u>his</u> engagement. He protests that the

person thus engaged is not the "real" he, thereby "ostracising a part of himself by regarding it as an alien force. On this model, self-deception is accomplished by means of self-amputation." (Abelson, 1977, p.98)

Having disavowed the engagement, the self-deceiver is then forced into protective, defensive tactics to account for the inconsistencies in his engagement in the world as acknowledged by him. Fingarette states that the self-deceiver develops a "cover-story". Initially, as he comes near the danger zone or "hidden area" in question, there will, during the normal spelling out of his engagements in the world, be "breaks" or "gaps" of non-spelling out. The covering up of his own refusal to spell out (i.e. he does not spell out why he is not spelling out) is secondary deception. Spelling out takes place because of the presence of certain reasons and non-spelling out takes place because there are reasons <u>against</u> doing so. It seems then that Fingarette's skill of spelling out has a daunting task in self-deception: not only must the skill be able to spell out reasons for <u>not</u> spelling out a certain engagement, it must also be able to cover the gaps that will appear in his normal spelling-outs, whenever he gets too close to the distressing engagement that is hidden from the rest of the self. In other words, the skill of spelling out must be able to spell out why it is not spelling out. Fingarette has to explain not only how the self-deceiver does not "spell out" his engagement, but also how he has somehow rendered himself incapable of doing it. The answer to this is that the skill in spelling out must be self-covering, i.e. it must enable the individual successfully to avoid spelling out its own exercise involved in assessing the reasons against the spelling out of the engagement in question. For otherwise, the whole exercise will be self-defeating.

> "I find there is a strong and preponderant reason for not spelling out — even to myself — that I have been a failure in realizing a certain ambition.

> Consequently I adopt the policy of not spelling this out. What is more, this policy obligates me as well not to spell-out my having made such an assessment of the situation and my having adopted this tactic. For, obviously, to spell-out my assessment would be to spell-out that I consider myself a failure, and that there are reasons for not admitting this even to myself ... And to spell-out the policy adopted as a result of the assessment would be to spell-out the fact that, though I have been a failure, my policy now is not to spell this out even to myself. In either case— whether I spell-out the assessment or the policy— it would amount to a clear abandonment of the would-be policy." (Fingarette, 1969, p.49)

Self-deception then consists in a self-covering exercise of the skill of spelling out in deliberately avoiding spelling out a particular engagement of the individual. The self-deceiver uses this skill "as inventively as possible in order to fill in plausibly the gaps created by his self-covering policy .... Out of this protective tactic emerge the masks, disguises, rationalizations and superficialities of self-deception in all its forms." (p.50).

Fingarette's theory has certain attractions. It seems to offer a solution to the epistemological paradox in that the availability of the skill of spelling out enables the individual not to be explicitly conscious of certain motives, aims and intentions of his, by refusing to spell these out. By the same token, the account seems to explain also the psychological paradox by laying bare the intentional character of self-deception; the shame in acknowledging the engagement as his motivates the individual to disavow the engagement in question. Furthermore, our feeling that there is in some sense deep insincerity involved in self-deception seems to be justified by the account.

But by circumventing the "cognition-perception" paradoxes, the account encounters paradoxes of a different kind. 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. He does say that the self-deceiver "persistently avoids spelling-out some feature of his engagement in the world" (p.47) but it is not obvious how to construe this analysis, especially because there is no explicit indication of what sort of proposition the self-deception is supposed to be about. The weak point of the theory is that the skill that Fingarette talks about has simply too much work to do. It seems as though the skill to spell out must spell out to itself what it must not spell out to the person, in other words, it must nullify its own exercise since the skill must not spell out. Fingarette's theory invokes another equally paradoxical notion: it seems to be the case that in self-deception one recognizes and yet deliberately avoids recognizing; one grasps and yet deliberately avoids grasping. Therefore, Fingarette's skill which is supposed to explain the paradox is itself paradoxical, insofar as it has to do, at one and the same time, two contradictory things: to spell out and not to spell out.

Moreover, the skill invokes the problem of consciousness. We have the paradox that the self-deceiver is engaged in the world somehow, he assesses this fact, but finds over-riding reasons not to spell it out. Either this engagement-plus-assessment-plus-decision operates entirely without consciousness or it does not. If it does operate without consciousness, then self-deception is purely figurative, a metaphor based on other-deception, but not literal self-deception. If it does not, then the self-deceiver must somehow be conscious of the engagement, or part of it, and the theory is returned to the paradox it wished to evade in the first place. Furthermore, the question arises whether the inability to spell out his engagement is an involuntary inability (rationally or psychological incompetent? or

prevented by some alien force?) or is it a voluntary refusal, a deliberate rejection? Does the self-deceiver really believe that the engagement is not his own or does he merely pretend so?

Haight argues that we cannot understand self-deception merely in action-volition terms: "as long as self-deceivers are people, 'know' and 'believe' must apply to them somehow." (1980, p.92). Of course, deception in general does have an action-volition side (A makes B believe that p), but it tells only half the story. It is not at all clear that Fingarette's skill of spelling out can operate except in terms of language and cognition.(7) The essential thing about the skill is that it enables the individual to assess the reason for and against spelling-out a certain engagement. How can one assess reasons without using concepts like: noticing, recognition, belief, argument, etc.?

Another objection can be raised. It is true that a person deceives himself through his own agency, but in deceiving himself a person is not only an agent, he is also the victim of his own act of deception. Fingarette's theory does not offer an adequate explanation of how the self-deceiver is also a passive victim, a state which marks the distinction between self-deception and hypocrisy. Moreover, Fingarette commits the error of interpreting all cases of self-deception as being motivated by shame. It certainly is a serious kind of self-deception, but definitely not the only kind. For example, there is no shame involved in the case of the mother who deceives herself about her son's death — distress, anxiety certainly, but no shame or guilt. She may be criticized for being irrational but not for non-acknowledgement of her moral duties. Abelson (1977, p.98) points out that even in cases of moral self-deception, Fingarette fails to distinguish between cases in which the self-deceiver excuses himself on the

grounds that his actions are not really his own (the only case Fingarette examines, although not the ones in which the agent pleads coercion by some force) and cases in which the self-deceiver has genuinely persuaded himself that he was justified in a course of action that by all objective standards is immoral.

And lastly, if the self-deceiver does not avow the engagement as his, does not accept responsibility for it, why then the elaborate "cover up" tactics? If he doesn't avow it as his in the first place, there is no need to employ self-deceptive strategies so that he does not avow the engagement as his. Fingarette's own example (1969, p.30) is of a person who has cancer but refuses to acknowledge it. The cancerous-self is, therefore, an engagement which he disavows, and avows only the healthy-self as engagement. The cancerous-self is, thus, no part of his synthesized identity, his belief-system, his values, his attitudes, etc. But if it is not there, why deceive himself about it? After all, the disavowed cancerous-self is not spelt-out, not "registered in the mind". Following Pears' argument about irrationality (see beginning of Chapter 3), if the conflicting belief or awareness is not registered in the mind, the question of irrationality is greatly deflated.

So it seems as though self-deception must include a "cognition-perception" element, but just how rigidly beliefs and thoughts are applied to the self-deceptive process may be an area worth investigating. It has been suggested that the self-deceiver intentionally misinterprets the evidence, and to do this he must believe that the evidence does count strongly against p, while both interpreting it not to count strongly against p and believing his interpretation. Mele (1983, p.372) suggests that this is not necessarily the case. The person's misinterpretation can be explained

on the weaker and unproblematic supposition that he recognizes or believes that the evidence might be taken to count strongly against p. Furthermore, to say that the person intentionally interprets the evidence in the way he does, is not to say that the person intentionally misinterprets the evidence, i.e. that it is his intention to misinterpret the evidence. It may be the case, of course, but it is not necessarily so. As Mele remarks, the self-deceived cancer patient believes and sincerely avows that he will not die of cancer; nevertheless, he may believe that the chance that he is wrong is significant enough to warrant looking into funeral arrangements.

At the beginning of this chapter, I referred to the sliding scale of cases of self-deception. In case (1) where there is balanced evidence for p and for not-p, it is obvious that the evidence allows enough latitude so that the agent can consciously choose the evidence he wants in order to support his belief that p. He is not guilty of forming an irrational belief in the face of strong counter-evidence. He is not even manipulating evidence to suit his belief or even avoiding evidence that supports the opposite belief. He is conscious of all the evidence, but he bases his belief on the evidence he wants. He is irrational only insofar as not suspending belief when faced with balanced evidence for p and not-p. As Pears (1974, p.103) notes, in case (1) there is no irrationality insofar as the agent does not flout the precept: "Accommodate your beliefs to your evidence." It flouts a different precept of reason: "Get all the available evidence that you need." This is irrationality of a different kind.

Davidson in How is Weakness of the Will Possible? develops the idea of latitude in a case of self-deceptive akrasia, in which the agent's premises do not actually entail the value-judgment that they indicate to be the

rational one to make about the particular predicament and so, under the influence of a rebellious desire, he is able to make the opposite value-judgment without contradicting himself. In this kind of case he has enough latitude to draw a consciously irrational conclusion. He makes a wishful belief on which he bases his value-judgment, and then goes on to perform the akratic action without inhibition. For example, the reformed smoker who accepts a cigarette knows that it might (but not that it will necessarily) revive the habit. He banks on the outcome that it won't and is, therefore, prepared to take the risk. In this case the self-deceiver knows that his belief conflicts with the evidence of studies done in the related field, but he also knows that the evidence is inconclusive. This latitude makes it possible for the wrong belief to be formed, i.e. "Smoking won't revive the habit" on which he bases his value-judgment that it is alright for him to accept the cigarette, which he does without inhibition.

I don't want to embark on a discussion of akrasia itself, but the idea of latitude is useful, because it can be applied to cases of self-deception. People very seldomly have absolute certainty that the available evidence is conclusive. It is common enough to hold onto a belief even though there is good counter-evidence, or to refuse to accept a belief despite good evidence for it, because evidence is often difficult to interpret and to judge. As Hanson remarks, "We must not forget the variety and the dimensions of human incapacities and failures." (1986, p.109)

Davidson describes the notion of latitude in the following way:

> "Probably it seldom happens that a person is certain that some proposition is true and also certain that the negation is true. A more common situation would be that the sum of the evidence available to the agent

> points to the truth of some proposition, which inclines the agent to believe it (makes him treat it as more likely to be true than not). This inclination (high subjective probability) causes him .... to seek, favour or emphasize the evidence for the falsity of the proposition, or to disregard the evidence for its truth. The agent then is more inclined than not to believe the negation of the original proposition, even though the totality of the evidence available to him does not support this attitude." (Davidson, 1985, p.139)

The analysis of self-deception operating with latitude is then that the self-deceiver holds the false belief that p on the basis that he believes that not-p is more likely to be true than p, and it is this thought which prompts him to behave in such a way as to cause himself to believe that p. He does this by means of the eight different strategies noted earlier on of one-sided evidence gathering and biased thinking. The paradox of self-deception is, thus, watered down to: self-deception is when the self-deceiver's motivation originates in the recognition that the evidence makes it more likely than not that not-p is true and that the self-deception is done with the intention of producing the belief that p, i.e. a belief in the negation of not-p, because the agent wishes that p.

At this stage I want to note an interesting solution presented by Priest to the epistemological paradox. This paradox rests on the assumption that the self-deceiver holds both the belief that p and its negation. But Priest notes that "to reject something is not to accept its negation. One can reject something without accepting its negation." (1986, p.105). This rejection of the belief that not-p does not involve the agent's stronger inclination to accept the negation of the original proposition (as with Davidson), but to suspend judgment altogether. For example, a man is extremely attracted to other men, is aware of this strong physical attraction, but because he does not want to acknowledge to himself his homosexual

tendencies, he deceives himself about them. But how is this self-deception instantiated? Instead of rejecting the belief "I have homosexual tendencies" and accepting its negation, i.e. "I do not have homosexual tendencies", he can remain agnostic. This means that he remains intellectually detached by neither believing it nor refusing to believe it. Since the self-deceiver holds neither the belief that p nor the belief that not-p, the epistemological paradox is dismissed. "It is, perhaps, the confusion between rejecting something and accepting its negation which is at the root of the view that one cannot believe a contradiction." (p.106). This view based on the trichotomy accept/reject/be agnostic (or what Parsons calls the assert/deny/neither trichotomy) certainly seems an attractive and plausible interpretation, but I think it is applicable only to peripheral cases of self-deception, rather than the typical and more problematic ones in which the self-deceiver does firmly hold a particular belief (rather than to suspend the holding of either belief) for which there is little or no direct evidence in "weak" self-deception, or strong, indusputable counter-evidence in "hard" self-deception.

The world is full of confusing facts and, according to the view of many people, conflicting evidence, e.g. the loving father who shoots his family; the manufacture of nuclear arms in order to maintain peace, etc. The self-deceiver has the comforting knowledge that inconclusive evidence allows him to select evidence that supports his favoured belief. Since his favoured belief is now "rationally" supported by evidence, the self-deceiver happily accepts the belief as being a rational belief. The knowledge that absolute certainty about the truth of anything is rare enables the person to adopt inconsistent beliefs. By invoking the notion of latitude, the meaning of "know" and "belief" is considerably watered down. "Know" is more like "suspect", and "belief" shades into "opinion". With the meaning

of the terms watered down, the paradox is, therefore considerably weakened.

> "(S)elf-deceivers neither just believe, nor both believe and disbelieve, nor just disbelieve what they present themselves as believing. Instead, they must be seen as people who are genuinely uncertain whether something that they want to believe is really so, and who decide to act as if it were so not in the hope of deceiving others, but in the hope of eliciting decisively confirming evidence from others. This project of 'experimentation by bluff' is a natural part of youth's adaptation to reality, and may well form the preceeding stage of what later becomes a properly self-deceptive reaction to reality." (Kipp, 1980, p.316)

Kipp's interpretation is certainly applicable to the "weak" case (1) of self-deception, but not to all cases. One of the features of a more radical type of self-deception is that the self-deceiver deliberately avoids "confirming evidence from others".

I stated that the notion of latitude weakens the paradox of p and not-p. The way in which latitude deflates the epistemological paradox is that it allows the agent consciously to form an irrational belief when his other beliefs (i.e. the belief that the evidence is inconclusive) allow him to form it without self-contradiction. Latitude reduces the irrationality of the belief formed in cases (1) and (2) since the evidence allows that belief to be formed, but the belief, however, still remains irrational because it is caused purely by a wish. The agent chooses to believe that p because he wants to believe that p. Naturally, the agent's latitude allowed by the evidence will diminish as his premises become stronger and make the opposite belief more probable, case (2), but it will not vanish altogether until they actually entail the opposite belief, cases (3) and (4). In these latter cases, the self-deceiver cannot consciously form a belief that does not fit a perceived fact in a case where there is no latitude. "For in such a case he would know which proposition fitted the fact and he would

have to believe the contradictory proposition and no sane person can believe this conjunction." (Pears, 1982b, p.174). In the next two chapters I shall deal specifically with these problematic cases.

As I have stated at the beginning of the chapter, these aforementioned solutions to the paradoxes of self-deception do give valid accounts of the phenomena we label as "self-deception", but according to Pears' sliding scale, these various solutions cannot account for cases (3) and (4). In these cases the self-deceiver cannot appeal to latitude since the conclusion is necessarily entailed by the premises or the self-deceiver cannot evade or distort the counter-evidence because it is too strong and overpowering for a sane person to avoid or change to suit his own desires. The above approaches were based on the definition of self-deception as persuading oneself to believe something which one suspects to be false. If, however, we adhere to Foss' condition for self-deception, we may well find ourselves returned to the paradoxes. "One thing not always recognized about deception is that knowledge is a prerequisite: to deceive another, one must bring it about that the other believes what one knows to be false." (Foss, 1980, p.241). Although this is not necessarily true of all deception, certainly not of "weak" deception, it does seem to apply to "hard" cases of deception, i.e. case (6) of other-deception and cases (3) and (4) of self-deception. It may seem as though I am contradicting myself by invoking Foss' requirement of knowledge for "hard" cases, when on p.112 I invoke the requirement of knowledge for intentional cases of self-deception. However, the interpretation of "knowledge" differs. Knowledge in intentional cases entails an awareness or at least suspicion about the truth of the belief that not-p (or about the falsity of the belief that p), whereas knowledge in "hard" cases entails a certainty about the truth of the belief that not-p. Add to this the Sartrean qualification

that the self-deceiver "must know the truth very exactly in order to conceal it more carefully", and we seem back in the quagmire of the epistemological paradox. The solutions presented in this chapter rely on a less strict definition of deception (i.e. to persuade oneself to believe something which one suspects to be false), whereas in "hard" cases the stricter definition is employed (i.e. to persuade oneself to believe something which one knows for certain to be false). In these cases, it does not seem possible to give a coherent account without invoking a Theory of Systems.

The criticism levelled at the interpretations of "weak" self-deception may be that these views regard that a person who has deceived himself about a particular matter as no more than that his judgment was mistaken and that he should have known better. However, while this is the case in "weak" self-deception, such uses appear to be peripheral and not to reflect the central features of the concept as normally understood. "Weak" self-deception essentially implies some kind of deliberate refusal to face the truth or to interpret the truth realistically. But surely matters are different when the self-deception is brought about and maintained by the self-deceiver's intentionally concealing from himself what he at the same time knows for certain to be the case, or when the counter-evidence is too obvious and overpowering to ignore, or when the deductive evidence necessarily entails the belief that not-p.

> "We seem, in other words, to have arrived at the following unsatisfactory position: either we must accept a watered-down version of the concept that fails adequately to capture its primary function in thought and language, or else we must simply acquiesce in the existence of paradoxes viewed by many philosophers as falling well below the threshold of what is logically tolerable." (Gardiner, 1970, pp.231-232)

In the next two chapters I want to show that the different Theories of Systems succeed in accounting for "hard" self-deception, and that these theories do not fall "below the threshold of what is logically tolerable".

* * * * *

In this chapter I looked at various solutions that tried to explain away the paradoxes of self-deception. These succeeded in doing so, by basing their interpretation of "deception" on the loose definition, an interpretation that allows the self-deceiver intentionally to cause his own self-deception without an inherent paradox. By appealing to latitude, either within the evidence which is inconclusive, or latitude within his own inductive reasoning process, the "weak" self-deceiver is able to form consciously two opposing beliefs. Other theories again have succeeded in explaining away the paradox by denying that self-deception necessarily entails an opposite belief. But whatever the theory, as I mentioned in the conclusion, a cardinal feature of "hard" self-deception seems to be lacking, i.e. the knowledge that the self-deceiver has not only of what is necessarily true, but also the knowledge that he is evading, distorting or suppressing this truth. However, there are various degrees of self-deception, not all cases of self-deception are "hard" cases, and the theories presented in this chapter adequately account for the phenomena of "weak" self-deception. I disagree with the view that these cases are "phenomena we falsely call 'self-deception'", since these "weak" cases of self-deception are analogous to "weak" cases of other-deception. Not <u>all</u> cases of self-deception fall within cases (3) and (4) and not all cases of other-deception entail a direct lie, case (6). But to leave the investigation into self-deception here is to tell only half the story. I shall

attempt in the next two chapters to present the other half, a half which does not exclude "weak" self-deception, but one which complements it to give a coherent account of self-deception in general.

Notes

1. In order to solve the apparent problem posed by conflicting desires, Plato divided the soul. However, there is nothing intrinsically paradoxical about consciously having incompatible desires. After all, this is an all too human affliction. We may even be aware that our desire cannot logically or physically be satisfiable, but that does not necessarily lead to the abandonment of that desire. However, the case of conflicting beliefs is different, because beliefs aim at the truth. The idea of two conflicting beliefs within one conscious belief system certainly seems paradoxical, but this is the case of "hard" self-deception with which I shall deal in Chapters 5 and 6.

2. See Fingarette and Elster (1979, p.179).

3. This is in direct opposition to Foss who states that knowledge is a necessary condition for deception. " .... to deceive another, one must bring it about that the other believes what one knows to be false." (1980, p.241). See also Haight (1980, p.9).

4. For the sake of consistency within this thesis, I have swopped both Bach's and Haight's notation of p and not-p. Throughout my work the belief that p refers to the favoured, but irrational belief——the self-deceiver desires that p and comes to believe that p, whereas not-p is the truth which generates the unwelcome belief that not-p. Bach has chosen the exact opposite notation, i.e. the belief that p is the unwelcome belief and the belief that not-p is the favoured belief. I have changed all his notations, even in the quotes, to conform with my choice of notation.

5. For an explanation of the mechanics which make mental distance possible, see the Freudian and Functionalists' Theories of Systems in Chapters 5 and 6 respectively.

6. Haight herself raises an objection to this interpretation of literal self-deception by A. A_1(the one "self") and A_2 (the other "self") may be in the same cerebral cavity, but clinical experiments have shown that these two selves have two different belief systems, personalities, memories about experiences, etc. It is, therefore, debatable whether A_1+A_2 form a "self", a psychological unity which deceives itself.

7. Other philsoophers have noted this objection and have developed Fingarette's theory to include terms of cognition. See Martin (1986, p.13ff): "Full and sincere acknowledgement to others entails knowing or believing what one is saying. But it goes beyond mere cognitive states by being a revelation or open expression of what is known .... There are some rough analogies between acknowledgments to others and to oneself. Both involve bringing into the open something that has been or is a likely candidate for concealment .... (S)elf-acknowledgment especially requires that the belief (or knowledge) becomes active by being brought into an intimate relation with other aspects of the personality. There must be a willingness to integrate it into one's conscious thoughts, reasoning, emotions, attitudes, values and actions." (p.14 - 15). Cf, also Elster's (1979, p.178ff) view of self-deception in terms of self-modification.

CHAPTER 5

THE FREUDIAN THEORY OF SYSTEMS

In the last chapter I looked at various theories pertaining to "weak" self-deception in which the characterising feature is the self-deceiver's appeal to latitude, either in the way in which the often ambivalent evidence can be interpreted or in the latitude allowed via inductive reasoning in drawing a conclusion. However, if we move onto cases (3) and (4) the latitude is severely curbed and the self-deceiver can no longer ignore the fact that he now holds two inconsistent beliefs, since the evidence which presents itself to him necessarily implies the conclusion that not-p. As I have mentioned earlier, two inconsistent beliefs within the same belief system point to the fact that the beliefs cannot all be true. Since no sane person can rationally believe that an explicit contradiction is true (i.e. aB(p+-p)), the initial belief and its explicit negation will have to be kept apart in the subject's mind (i.e. aBp+aB-p). Pears states that "The only thing that cannot happen is that two contradictory beliefs should be caught in the same focal consciousness and both survive. It really is impossible to believe the conjunction of two contradictory propositions .... if the cautionary belief is in the person's consciousness." (1984, p.76).

It seems as though the two inconsistent beliefs will have to be assigned to different systems within the same brain if successful "hard" self-deception is to be achieved and maintained. I have looked at how mental distance and rôle playing can, to a certain extent, succeed in maintaining the chasm, but the more irrational the belief or the more overwhelming the counter-evidence, the more drastic the self-deceptive tactics the agent must employ. Mental distance, Orwellian "doublethink" or self-deceptive rôle

playing can no longer do the job when the self-deceiver clearly understands the concepts he is using. Then aBp+aB-p is possible as long as we grant the possibility of two belief systems within the person. The different theories of Systems have different interpretations as to what these two systems encompass or how they are brought about (i.e. the criterion for the division). In this chapter I shall look specifically at the Freudian theory of the Divided Mind and at Pears' and Fingarette's modifications of it. In Chapter 6 Davidson's Functional theory of Systems will be examined. The aim of the examination is to see how these theories overcome the four paradoxes of self-deception.

My aim in this chapter is not to give a psychologically detailed account of Freud's theory, but to sketch it broadly in general terms in order to examine it from a philosophical point of view. Freud's theory about the way in which the two contradictory beliefs are kept apart in the self-deceiver's mind is that the one that gives satisfaction remains in consciousness, while the other is kept out of consciousness. The criterion for the division in Freud's theory of the Divided Mind is, therefore, that of consciousness. One of the two contradictory beliefs can be held consciously in the sense that the self-deceiver is readily able and willing to attend to it, whereas the threatening belief is placed in the unconscious in the sense that the self-deceiver cannot readily attend to it (except when he is under special influences like hypnosis, drugs, psychotherapy, etc.). The unconscious belief which is suppressed (the unwelcome true belief that not-p) is the basis for the self-deception.

Freud's view has its attractions: if someone forms a very irrational belief, it seems obvious that he cannot consciously believe that it is very irrational. Freud was concerned mainly with the unconscious, which

is a permanent reservoir of deeply suppressed wishes, drives and thoughts. However, as I have illustrated in Chapter 4, in most cases of self-deception there is no deep suppression into the unconscious but rather a shifting away of attention from the rational belief, or only a shallow suppression into the preconscious (in the sense that the self-deceiver is not readily able, but is able with a certain amount of effort to attend to it), and even that only needs to last as long as is necessary for the particular piece of self-deception. I do not want to embark on a discussion of how the theory of the Divided Mind can be used to explain "weak" self-deception, which would entail suppressing the belief that not-p into the preconscious —— this is like evading not-p because one has a suspicion that not-p, but this suspicion is never clearly and consciously examined or evaluated by the agent. Invoking the theory of the Divided Mind with all its problematic philosophical implications in order to explain "weak" self-deception is like killing an ant with a sledgehammer. It can be done, very effectively, but the same result can be achieved by less drastic means. The "less drastic means", in this case, occur in the theories discussed in Chapter 4. This chapter will concentrate on the theory of Systems as applied to "hard" self-deception.

The word "conscious" forms the core of the Freudian theory and it is, therefore, imperative that the meaning of the word is defined. It is exactly this area which has given rise to various confusions, since there are at least two meanings of "consciousness", with some interpretations of the Freudian theory not stipulating which meaning is used, or even with the meanings used interchangeably.(1) One meaning of "consciousness" is the name of the main system which controls a person's life and the other meaning of "consciousness" is the name of a relation between any system and the

elements to which it reacts, a "being aware" of certain elements even when they belong to a different system within the person. Traditionally, Freud's split in the psyche was conceived to be between the Conscious and the Unconscious, each eventually conceived as a system. The two systems are quasi-autonomous, often incompatible with and alienated from each other. The Unconscious is an illogical cess-pool which operates according to "primary processes" in the absence of causal and temporal relations, whereas the Conscious operates according to the more rational "secondary processes". The two systems interact by way of conflict rather than co-operation. The central problem of the theory of consciousness as a system is to offer an explanation of how the rational and cautionary beliefs are pushed into this illogical system, whereas the irrational belief is allowed to flourish in the more rational system.

The later Freud describes some kinds of irrationality as forms of self-division in which consciousness splits itself into different systems. The "conscious" to which the later Freud refers should be interpreted in a more "functional" sense, according to Pears (1984, p.37), Fingarette (1969, p.130) and Clavell (1986, p.504). In the "functional" sense, the line between various systems is drawn to reflect the interactions between a person's mental attitudes.

> " .... in one and the same individual there can be several mental groupings, which can remain more or less independent of one another, which can 'know nothing' of one another and which can alternate with one another in their hold upon consciousness .... If, where a splitting of the personality such as this has occured, consciousness remains attached regularly to one of the states, we call it the conscious mental state and the other, which is detached from it, the unconsciousness one." (Freud, "Five Lectures on Psychoanalysis" S.E. vol.XI, p.19)

Since these various groupings can "alternate with one another on their hold upon consciousness", the name "conscious" is not the fixed name of a fixed, static system, but is rather a label that can be attached to a particular mental state, depending on the level of awareness of that particular mental state. The functional meaning of "conscious", therefore, refers to a dynamic quality which can be attached to different systems at different times, and to the dynamic interaction that exists between different systems. Consciousness is not the name of a static pool in which certain ideas are entertained, but consciousness is itself an active mental force. The functional sense of consciousness means that different systems or elements come to attention (a kind of inner eye) so that they are conscious when attended to. Consciousness, therefore, is interpreted as being active, rather than passive.

I shall first look at Freud's theory with "conscious" referring to a passive separate system and later at Pears' and Fingarette's interpretations of the Freudian theory with "conscious" referring to the active relationship that holds between the system and the elements to which it reacts. It is especially this functional interpretation which will form the basis for Davidson's theory. (see Chapter 6).

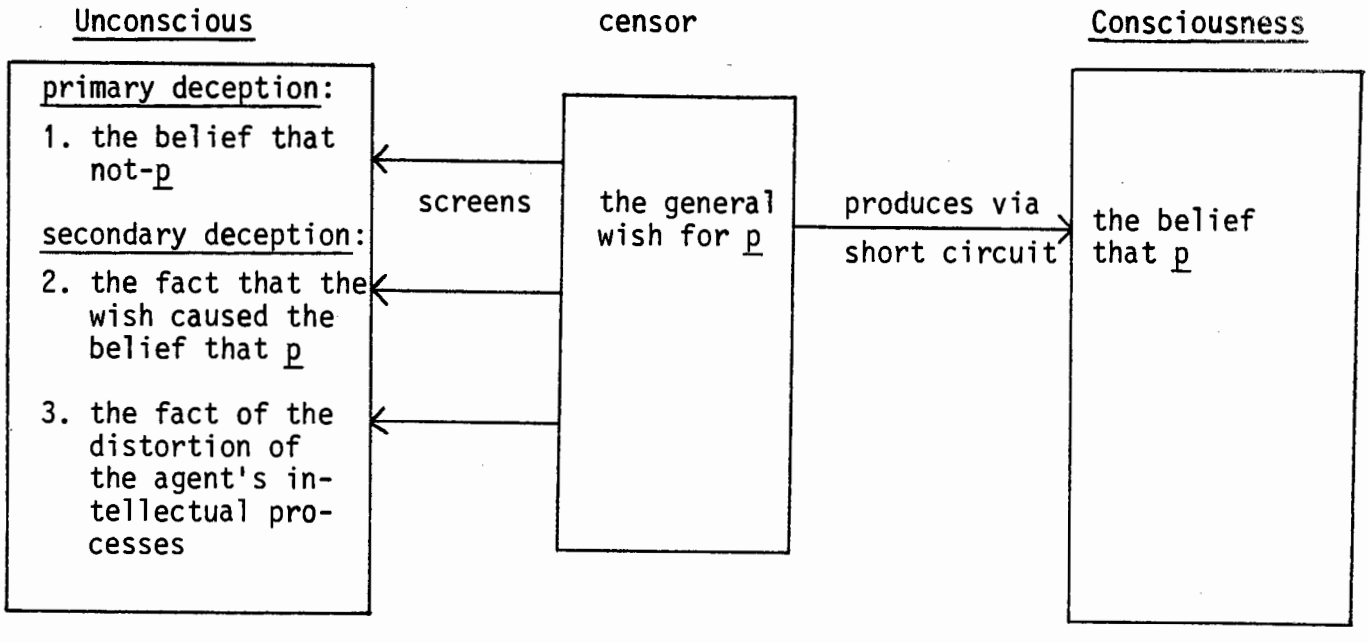
> "(S)ome mental phenomena that we normally assume to be conscious, or at least available to consciousness, are not conscious, and can become accessible only with difficulty, if at all. In most functional respects, these unconscious mental states and events are like conscious beliefs, memories, desires, wishes and fears." (Davidson, 1982, p.291)

The early theory of the Divided Mind claims that the psyche of an individual is split, that there is some kind of duality within the mind. The conscious part and the unconscious part are kept separate by a "censor" which

prevents the self-deceiver (the self-deceiver does not prevent himself) from recognizing certain things about himself, e.g. deep-rooted fears and wishes. This is analogous to the deceived in a context of other-deception in that the deceived is prevented from learning about matters which the agent does not wish the victim to know. I shall start with Pears' interpretation of Freud which shows that paradox 1 (and 2, 3 and 4, for that matter) need not arise, i.e. "If I have deceived myself that p, I believe that p but at the same time I really know that not-p."

In Chapter 2 I showed that wishful thinking is quite possible, even though there is no evidence for p or even when there is evidence against p. What happens, via the Freudian short-circuit, is that the wish for p gives rise to the wish to believe that p which, in turn, causes the belief that p (when p is too difficult to bring about in the real world). The wish, therefore, produces 1) satisfaction that p when p is believed and 2) the actual belief that p. This Freudian short-circuit can be applied in self-deception as well. In the Freudian account, the general wish for p is the agent which sets the complex short-circuit in motion, initiating the whole process of self-deception. The central principle of this complex structure is the wish for p which is the suppressive agent, or the censor which stands guard in-between consciousness and the unconscious. Sartre conceives the psychoanalytical censor as "a line of demarcation with customs, passport division, currency control, etc., to re-establish the duality of the deceiver and the deceived." (1958, p.50). The duty of the censor is to screen various aspects from consciousness. First of all, it has to screen the rational but unwelcome belief that not-p, since two incompatible beliefs cannot both survive in the same conscious belief system. Only the belief that p is left in consciousness because it is this belief that gives satisfaction and the general wish of the self-deceiver is to avoid

stress and conflict. The censor, by screening the belief that not-p, is maintaining the state of primary deception, i.e. deception about the true state of affairs. Furthermore, the censor must also screen from consciousness the belief that he is being irrational, i.e. the fact that the wish itself caused the belief that p. This belief cannot be in consciousness for then the irrationality of the favoured belief will be immediately apparent to the rational person, and the favoured belief will collapse. Since the whole purpose of self-deception is to form and maintain the favoured belief, this cautionary belief that the favoured belief is caused purely by a wish and is, therefore, irrational must be screened from consciousness. The censor, on this level, is responsible for maintaining the state of secondary deception, i.e. to screen the belief that he is deceiving himself. Finally, the censor has to screen from consciousness the fact of the consequential distortion in the self-deceiver's intellectual processes which allows him to practise self-deception. This too is on the level of secondary deception. The agent can, of course, be conscious of the general wish for p, but if the self-deception is not to be dissipated, the censor must screen these three aspects from consciousness. There is nothing intrinsically irrational about the wish; it, however, becomes an agent of irrationality when it starts distorting reason. Diagrammatically, the view can be represented as follows:



Since the unconscious is inaccessible and, therefore, also the belief that not-p, it is clear how the above theory overcomes the paradox of how the self-deceiver can believe that p while at the same time knowing or believing that not-p. At first glance this seems like a neat package in which the knowledge that not-p as well as the knowledge that he is deceiving himself is safely screened from consciousness and the belief that p is the only one the self-deceiver is conscious of.

Before looking at the philosophical implications of this theory, I want to apply it to a hard case of self-deception in which the evidence allows no latitude. Emma is an unmarried devout Catholic who has given birth to an illegitimate child. Having had a child out of wedlock is for the pious Emma a particularly abhorrent sin which fills her with great anxiety. As a rational being, she deduces from the medical reports, her actual experience of the pregnancy and the birth, the arrangements she had to make for the adoption, etc., that she has had an illegitimate child and holds the corresponding belief. On the other hand, Emma's speech and actions deny the existence of the child or that she was ever pregnant. She never refers to the child or the confinement, has destroyed all cards and documents relating to the birth, got rid of the clothes given to her for the baby and does not acknowledge its existence in any way. Her speech and actions confirm her sincerely held belief, "I do not have an illegitimate child." Since it is impossible for Emma to change the actual state of affairs, the only way to get rid of the anxiety and dissonance caused by the conflicting beliefs is to revert to self-deception which is initiated by her strong desire to appear virtuous and to avoid shame, guilt and anxiety. This case has been greatly simplified, in that in typical cases of "hard" self-deception it is not merely a single belief, i.e. Emma's belief that she does have an illegitimate child, that has been suppressed, but a complex of beliefs and

feelings which, in turn, generate other conscious beliefs and feelings over a long time. How then is paradox (1) avoided in the case of Emma? First of all, her desire to appear virtuous and to avoid shame generates the belief that p, i.e. that she does not have an illegitimate child. This belief is based purely on her wish and for this reason Emma cannot be aware that the contribution of the wish was necessary for the formation of her irrational belief, otherwise the plan of self-deception will be defeated. Apart from screening the rational belief that not-p from Emma's consciousness (primary deception), the general wish to avoid shame and anxiety also screens its own contribution in the formation of the irrational belief, as well as screening from her consciousness the very fact that screenings of irrational processes are taking place (secondary deception). Emma may, of course, be quite aware of the existence of the general wish itself, but not of the irrational sequences of that wish if the self-deception is to be successful.

Pears applies the Freudian theory to special cases of self-deception by appealing to the notion of time in order to overcome the paradox of how a person can consciously decide to deceive himself. He states that "Sartre goes too far when he argues that we cannot appeal to lapse of time in order to remove the incoherence." (1974, p.106).(2) Self-deception is not necessarily like lying. There can be a time lapse between believing that not-p and wanting to believe that p and forming the belief that p. Audi remarks that "acts of self-deception are not acts of putting oneself into self-deception at a stroke; they are acts manifesting it or conducive to producing it, such as putting certain kinds of evidence out of mind, or sincerely denying something one unconsciously knows to be true." (Audi, 1982, p.142)(my emphasis). Self-deception, therefore, does not necessarily entail that the self-deceiver must believe that p and believe that not-p at

the same time. The self-deceiver need not hold these two incompatible beliefs either simultaneously or hold both in consciousness. He may decide with the help of the censor to get himself to believe that p when he believes that not-p. So, at the outset he does not believe that p, but during the process he relies on his general desire for p to gradually suppress the belief that not-p, or it weakens through the subject's avoidance of supportive evidence for not-p. Trusting to luck, the self-deceiver eventually consciously believes that p. However, the two beliefs are never simultaneously entertained in the same consciousness. The rational belief is either suppressed or it has faded completely. Not only does time overcome the inconsistency of the two beliefs (paradox 1), but Pears notes that with time the wish screens also the uncharacteristic distortion of the intellectual processes when the self-deceiver first knowingly embarked on his plan of self-deception, and so with time the self-deceiver becomes also less and less aware of the motive for the plan of self-deception in the first place (paradox 2). Pears adds that the plan of self-deception "cannot be fully reviewed when it approaches completion." (1974, p.107). It is the plan of the censor that with the lapse of time the motive —— the rational tendency to believe that not-p ——will be screened. Eventually the motive may come through in a different version, a version that will not contain the tendency to believe that not-p.

In order to remove the appearance of incoherence from Emma's plan to instill in herself a belief which is in direct contrast to a belief she already holds (paradox 2), Pears suggests that she, though quite aware of her rational tendency to believe that she does have an illegitimate child, can rely on her wish to believe that she does not. She can, at first, consciously and intentionally embark on the project of self-deception with the explicit intention of not dwelling on the distressing subject. She may include her

motive when describing her action in the following way: "I am going to dwell on my virtuous nature and disregard any evidence that may point to my lack of virtue because I suspect that I have sinned and I want to believe that I am virtuous." In other words, the first stage in the plan is to screen the rational tendency to believe that she has an illegitimate child. During the initial stages of the plan she is able to review her plan to get herself to believe that she has not sinned. With time her motive of the rational tendency to believe otherwise is screened from her consciousness, or if she is conscious of her motive it is in a way which does not include in any way knowledge of her sinful lapse. Emma's motive is, of course, an essential part of the whole process of self-deception at first, but there must come a stage where she can no longer be aware of her initial motive. Just as a time lapse brought about the fading from consciousness of the belief that she has sinned, so a time lapse can bring about the fading from consciousness of the intention to form the plan of self-deception. At this later stage, she may describe what she is doing, but her description will make no reference to her motives. The problem raised is that when Emma starts planning her campaign, she must make provision for this stage and has to decide in advance how she is going to proceed from this point on in the execution of her plan. In other words, she is faced with an extra task if her plan is to be successful. However, Emma does not know when this stage will appear and cannot plan for further action after this stage has passed. She has to rely on the "discreet operation" of her wish to believe in her virtuousness and to avoid shame, and it is reasonable for her to expect that this wish will guide her actions in such a way as to further her plan. Pears notes that the self-deceiver rarely plans "the strategy of his campaign in advance. What usually happens is that his project seems to improvise itself as it proceeds." (p.109).
In the context of other-deception, the deceiver's motive for saying that p

is that all the time he believes that not-p. Therefore, Emma's project of getting herself to believe that she is virtuous is similarly motivated by the belief that she has sinned. The original belief that not-p has either become weaker or it has faded from consciousness altogether. The belief that not-p can be weakened by using the tactics employed in "weak" self-deception (e.g. avoiding the photo albums in which there is undisputable evidence of her pregnancy, etc.) or the use of these tactics can have the result that the belief that not-p is still strong, but has gradually been screened from consciousness. Time, therefore, removes the belief that not-p as well as the motive for the plan of self-deception from Emma's consciousness. In this example of self-deception, thus, paradox 1 and 2 have been overcome.

But paradox 3 and 4 still have to be dealt with. Paradox 3 states that if the self-deceiver divorces his belief that not-p from the rest of his thoughts and beliefs, the process becomes unintelligible. In order to motivate the process of self-deception and to guide its strategy, an awareness of the belief that not-p is needed. And since paradox 4 (which states that not only the belief but the whole plan is incoherent and cannot be made possible merely by being screened from consciousness) is an extension of paradox 3, Pears maintains that the same arguments are applicable in showing how in Emma's case the paradoxes are overcome. The paradoxes are based on the assumption that the screen deprives the rational tendency to believe that she has sinned from its motivating force when, in fact, it keeps on motivating Emma in her plan of self-deception even though she is not conscious of the motive. One can argue that since the belief that she has an illegitimate child and has, therefore, sinned (this is, of course, not a necessary conclusion, but this is Emma's assessment of her own situation) is no longer conscious, Emma has no need to perpetuate the self-deception any longer.

However, even though the belief in her virtuousness has taken over from the belief that she has sinned, it does not mean that the motivating force of the latter belief will die away. The powerful motivating force of unconscious desires and thoughts is often demonstrated in cases of post-hypnotic suggestions and Freudian cases of hysteria. The rational tendency to believe that she has sinned is divorced from the rest of her conscious thoughts and beliefs, but it still guides and motivates Emma's actions and plan from behind the screen.(3) I shall return to the objections raised to this view a little later on.

There are various philosophical implications and problems in this interpretation of the Freudian theory of systems. First of all, the theory is based on the general wish which acts as a censor or screen. Looking at the diagram, it becomes apparent that the general wish for p must screen the fact that it caused the belief that p, in other words, the wish must screen an aspect from itself. Screening this aspect from the consciousness of the self-deceiver seems to call for a second wish, the wish to maintain an unawareness of the necessity of the contribution of the first wish in the formation of the irrational belief that p and a second (or perhaps, third?) wish to screen the consequential distortion of the self-deceiver's intellectual processes caused by the first wish. This objection can be extended to include an infinite regress of screening wishes, i.e. a third wish is needed to screen the second wish from consciousness, etc. Pears overrules this objection of an infinite regress by suggesting that the general wish practises self-reflexive suppression and need not be suppressed by another wish which, in turn, is suppressed by yet another, etc. Pears suggests that the unawareness of the necessity of the first wish in the formation of the irrational belief is not screened by a second wish, but is a further intellectual distortion produced by the first wish itself. He

interprets the wish rather as a single wish multiplying distorted intellectual processes, rather than a whole series of multiplying wishes behind the screen. So, instead of having many different wishes, each screening the previous one, we can postulate a general wish which will include, in Emma's case, the wish to avoid shame, to avoid conflict and anxiety, to avoid awareness of the irrationality, to avoid awareness of that irrationality and its contribution to it, etc. It is, therefore, possible, to explain the secondary effects of the screening without postulating a complex plan behind the screen. With this approach, Pears is "multiplying the effects of a single wish" instead of "multiplying wishes behind the screen." (Pears, 1974, p.101). However, exactly how the wish suppresses itself and gives rise to the consequential distortions is unclear, because we have no direct access to the mind of the self-deceiver, and even if we were ourselves deceived, we would have no direct access to the wish which has screened its operation from our consciousness. "The dynamism of Freud's theory exposes it to a dilemma. If the suppressing agency is always a wish, then either there is an infinite regress of suppressing agencies or suppression is self-reflexive. The second option is clearly preferable even if we have no picture of the way in which this self-reflexiveness works." (Pears, 1982a, p.275-6)

Another problem seems to present itself: surely it is irrational for a rational person to keep his own truth seeking faculty in check? How is this possible? Pears notes that it can be rational to curb the truth seeking faculty when investigating oneself. If we take the case of Emma; she could be rational in wanting to believe that she is virtuous and thus to believe that she does not have an illegitimate child, because if she were to confront consciously the belief that she does have an illegitimate child, that knowledge would shatter the self-confidence she so desperately

needs to build up her self-esteem. Not only could it be rational to screen her primary deception from herself, but her secondary deception as well, for the same reason.

Sartre objects that the difficulty that Freud's theory seems to remove merely re-emerges at a different level. The question Sartre asks focuses on the status of the censor itself. In the diagram the censor stands inbetween the unconscious and the conscious. However, if the censor is to perform the functions assigned to it — selecting the drives that need to be concealed, interpreting and resisting the analyst's probings, etc — it must presumeably be aware of the material in question and aware of its own activity in seeking to stop this material from becoming explicitly conscious. But now the paradox recurs within the censor, for it both knows but conceals certain aspects from itself what it knows. The question that Sartre asks is whether the censor is conscious or unconscious. He poses the problem in the form of a dilemma: if the censor is in the conscious, it must be aware of the irrational sequence and its necessary defeat by reason. If the censor is aware of the illegitimacy of the favoured belief, the paradox of self-deception will remain unexplained. However, if the censor is in the unconscious, Sartre states that it then would not be aware of what needs to be concealed. "I must know the truth very exactly <u>in order</u> to conceal it more carefully." (Sartre, 1958, p.49). According to Sartre, the translucency of consciousness demands that "(t)hat which affects itself with bad faith must be conscious (of) its bad faith since the being of consciousness is consciousness of being." (p.49). In other words, the censor must have complete understanding of the deception as well as of the truth it is altering. If the censor were unconscious, it will not know what the unwelcome belief is. How can the censor screen certain aspects from itself if it does not know what these aspects are? "All knowing is consciousness

of knowing .... But what type of self-consciousness can the censor have? It must be consciousness (of) being conscious of the drive to be repressed, but precisely in order not to be conscious of it." (p.53). If the deceiving agency lacks awareness of the illegitimacy of the desired belief, then we shall have to say that it lacks the awareness because it has screened certain facts from itself. This, of course, leads to an infinite regress of deceiving agencies. However, to say that the censor is in the unconscious is the uninteresting part of Sartre's dilemma, because nobody would suggest that the rebellious centre of activity deceives itself in the unconscious. If, on the other hand, we hold that the deceiving agency or censor is aware of the illegitimacy of the desired belief, we still sit with the paradox that was meant to be solved in the first place. The paradox arises from the requirement that throughout the process of acquiring the desired belief, the deceiving agency must maintain the opposite belief. In self-deception the censor seems to adopt a complex strategy and in order to put this complex strategy into action, the censor must have some appreciation of the existence and strength of the belief which is the negation of the desired belief. Sartre argues that since the deceiving agency is capable of planning the complex process of self-deception, it is rational with knowledge of the conscious and it must, therefore, reach and accept the rational conclusion that it is practising self-deception. As a rational faculty, the wish then must annul the process of self-deception. Even if it is said that the censor keeps only the embarrassing facts out of consciousness and does not necessarily keep itself out of consciousness, the problem will still not be solved. It offers no explanation of self-deception, but is simply a redescription of the as yet unexplained phenomenon. There is still the problem of how two systems, mostly independent of each other, manage to coexist in a single person.

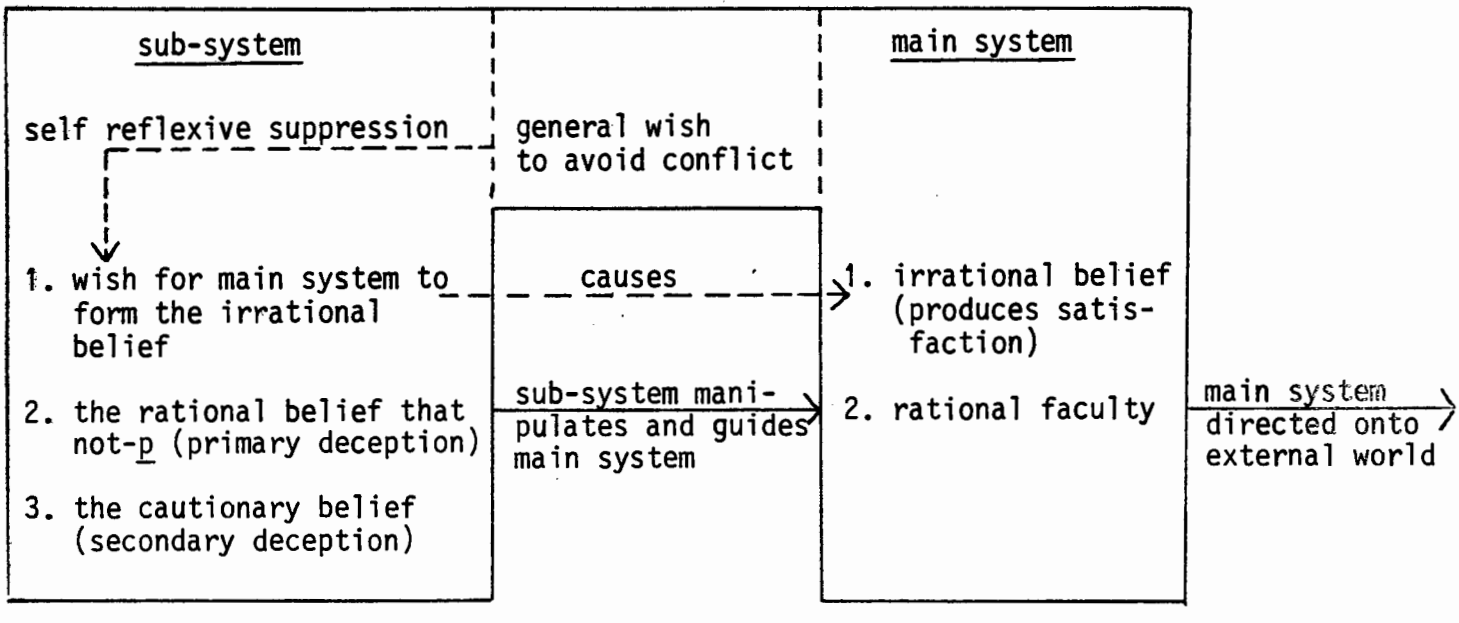
> "The very essence of the reflexive idea of hiding something from oneself implies the unity of one and the same psychic mechanism and consequently a double activity in the heart of unity, tending on the one hand to maintain and locate the thing to be concealed and on the other hand to repress and disguise it. Each of the two aspects of this activity is complementary to the other; that is, it implies the other in its being. By separating consciousness from the unconscious by means of the censor, psychoanalysis has not succeeded in dissociating the two phases of the act." (Sartre, 1985, p.53)

Pears extends Freud's theory and replies to Sartre's criticism as follows:

> "This is not a convincing argument. The natural objection to it is that the consciousness out of which the censor has to keep the unwanted belief is only the main system of desires and beliefs which run the daily life of the person. There is no reason why the censor itself should not be conscious of the existence of the unwanted belief in the unconscious. This would be entirely compatible with its consciousness of the existence of the opposite belief in consciousness. In fact, the censor might even share the unwanted belief with the unconscious." (Pears, 1984, p.37)

However, this solution clearly calls on the second meaning of "conscious", i.e. the functional <u>relationship</u> that holds between the two systems, now referred to as the main system and the sub-system. The main system controls the agent's daily life and includes the favoured but irrational belief as well as the information that makes it irrational. This is the system with the alleged paradox for it contains both the irrational belief and the rational faculty of the person. However, the inability of the rational faculty to recognize the favoured belief as irrational is due to the sub-system which includes "the cautionary belief, that, given his information, it was irrational to form the favoured belief". (Pears, 1984, p.67). This is, of course, on the level of secondary deception in that the belief that the belief in the main system is irrational cannot be in

the main system and the general wish to avoid conflict banishes the cautionary belief to the sub-system. Since I am concentrating on "hot" cases of self-deception the agent's wish is the culprit which upsets his reasoning. The rebellious desire to believe that p (and its contribution in the distortion of the intellectual processes) is aware that it cannot remain in the main system otherwise the main system's rational tendency will override it. The desire's only course of action to ensure the survival of its resultant irrational belief that p in the main system, is to take awareness of its own necessary contribution in the formation of the belief that p out of the main system and for the desire to set up and act as nucleus for the sub-system. The irrational belief that p remains in the main system but since the cautionary belief is no longer in the main system, the rational faculty of the main system fails to recognize the irrationality of the irrational belief and, therefore, does not interfere with it. The sub-system exploits the main system to allow both the rational faculty and the irrational belief to survive side by side. Diagrammatically, the systems can be represented as follows:



This diagram differs from the previous one one a number of issues. First of all, the Sartrean citicism of whether the wish is conscious or not is

no longer applicable. The general wish practises self-reflexive suppression on part of itself, viz. the rebellious wish. The agent can be aware of the other aspects of the general wish. So, instead of belonging either to the unconscious or to the conscious, the general wish in this later interpretation belongs to both systems. Furthermore, the later theory places far more emphasis on the dynamic and functional aspect of the two systems. Instead of a wish separating the two systems, now the sub-system directly influences the main system. But the most important difference between the two interpretations is that the sub-system comes into being only when there is a need to conceal irrationality from the agent's main system. Once there is no more need for self-deception, the consciousness-preventing relationship falls away. In contrast, Freud's earlier theory postulated a permanent, unconscious well whether the agent was being irrational or not.

However, this later interpretation also has some problematic philosophical implications. The principle on which it rests is that "if a person consciously believed that the belief he was forming was irrational, that would prevent him from forming it, whereas, if the same cautionary belief were not conscious, it would not prevent him from forming it." (Pears, 1984, p.72) In other words, for Freud the permissive cause of irrationality is a failure of consciousness. Pears notes that this principle can be divided into two parts. The first part then: "If a person's cautionary belief, that the belief that he was forming was irrational, were conscious, it would prevent him from forming the irrational belief." Therefore, if the agent detects the operation of the wish and recognizes its results as irrational, he will stop deceiving himself. However, this does not always happen. Very often the wish in consciousness, and its consequential distortions of the intellectual processes, may "still retain some of its power to fascinate

and delude." (p.73). The consciousness that a belief is caused purely by a wish is not sufficient to prevent it from being formed. Pears extends the theory further down the scale to milder cases of irrationality and it is here that he comes to the conclusion that "it is quite certain that mild cases are not always stopped by consciousness." (p.74). However, in this chapter I am concerned with whether "hard" cases of irrationalities are stopped by consciousness of their nature and causation. Applying the principle to "hard" cases, Pears concludes that it is "plausible to suppose that serious cases are always stopped." (p.73).

The first part of the principle, i.e. the non-permissiveness of a conscious cautionary belief, holds in serious cases of self-deception, but what about the second part? The second part of the principle addresses the question of the permissiveness of an unconscious cautionary belief —— "if the same cautionary belief were kept out of consciousness, it would not prevent him from forming the irrational belief." (p.78). This is an important point, for there would be no point in banishing the cautionary belief from the consciousness of the main system if it still obstructed the formation of the irrational belief when it (i.e. the cautionary belief) is in the sub-system. In other words, it must lose its power to obstruct when it is banished from the main system.

I shall now return to the objection I raised earlier on: Emma's rational tendency to believe that she has sinned is divorced from the rest of her conscious thoughts and beliefs, but it still guides and motivates her actions and plan from behind the screen. If the second part of Freud's principle is applied then, as Emma with time becomes less and less aware of the motive for her self-deception, as well as of how she is to fulfil her plan (she was unable to formulate the plan in its entirety from the outset), the plan

of self-deception will become less and less effective and will eventually fail for it is that very motive which steers the plan of self-deception in the desired direction. The motive behind the project will, therefore, gradually lose its force as the project reaches completion. However, underneath Emma's increasing belief that p, there is still the screened but undiminished rational tendency to believe that not-p. Pears states that "contrary to Sartre's assumption, the fact that it is screened from you will not necessarily deprive it of its motivating force. Your wish to believe p emerges as the plan to deceive yourself that p only because you have a rational tendency to believe not-p. The fact that you have this tendency may continue to produce its effect even when it has been screened from you." (Pears, 1974, p.107)

Therefore, it is not clear whether unconscious beliefs are, in fact, powerless to produce their normal effects in consciousness. Certainly, our actions are often controlled by preconscious beliefs. For example, Emma might avoid the photograph album containing evidence of both her pregnancy and of her illegitimate child because she suspects she might find distressing pictures there, and yet she might not be conscious of this belief. But there is a twofold problem here:

1. her irrational belief that she does not have an illegitimate child is explained by keeping it in the consciousness of the main system, whereas the cautionary belief is kept out of the consciousness of the main system and is banished to the sub-system.

2. However, her avoidance of the photograph albums is governed by a preconscious belief (i.e. a suspicion that the albums may contain distressing evidence), one that must be producing its normal effect in consciousness.

How can the belief in (2) produce its normal effect in the consciousness of

the main systems, but the cautionary belief in (1) does not —— if it did the self-deception would cease. In order to address this problem it is necessary to look at the structure of the Freudian sub-systems and the "key to the problem will turn out to be their rationality. They are built around the nucleus of the wish to believe and this wish is the cause that blocks the normal effect of the cautionary belief." (Pears, 1984, p.80/81). In other words, because the sub-system is built around the rebellious wish, it acts rationally when it prevents the cautionary belief from producing its normal effect in consciousness but allows the girl's preconscious beliefs about the photograph album to produce its normal effect, by diverting her attention to picking up a novel. The sub-system is, therefore, a dynamic and rational system. It is on this point that Freud's and Davidson's theories both agree.

> "The sub-system is built around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate centre of agency within the whole person, it is, from its own point of view, entirely rational. It wants the main system to form the irrational belief and it is aware that it will not form it, if the cautionary belief is allowed to intervene. So with perfect rationality it stops its intervention." (Pears, 1984, p.87).

So, for successful self-deception to take place the sub-system must be able to dominate and guide the main system (e.g. to avoid Emma from paging through the albums). From the above discussion, it is seen that a condition for the manipulation of the main system by the sub-system is the internal rationality of the latter system. However, there are further conditions which apply to the sub-system. Apart from 1) being internally rational, the sub-system also 2) has to react to certain elements in the main system in order to manipulate it successfully (if the sub-system were totally divorced from the main system it could not exert any control over it),

and the sub-system must also 3) be aware of the main system's problem which is that there is evidence pointing to an unwelcome conclusion. Furthermore, it 4) ought to be aware of the achievement of its goal, viz. the main system's eventual formation and maintenance of the desired belief. These additional conditions seem to threaten the internal rationality of the sub-system which has to acquire the information without acquiring the elements themselves. The cautionary belief in the sub-system makes it impossible for the sub-system to accept the evidence from the main system and to draw its wishful conclusion, nevertheless. Surely, if the sub-system contains the main system's evidence as well as the cautionary belief, it must form the rational belief that she has had an illegitimate child. This the sub-system does do. But a discrepancy is immediately apparent: how can the sub-system hold this belief if it is built around the wish to believe the opposite? "Whatever the sub-system, and however it is marked off from the main system, it cannot accommodate the rational belief, if it is built around the wish to form the irrational belief. In fact, it may not even be able to accommodate the evidence and the cautionary belief, because, together, they push it so hard towards the rational belief." (p.90).

Two complementary solutions are applicable to the above problem. The first one is based on what I shall term Pears' Principle: "A system can react to a belief or desire in another system without necessarily sharing it." (p.88). So, whenever the sub-system reacts to a belief belonging to the main system, it does not necessarily mean that the sub-system must contain that belief. The sub-system can, therefore, become aware of the evidence in the main system without taking in that evidence into its own system, for if it did, the sub-system would lose its own internal rationality, a necessary condition for "hard" self-deception. The second solution focuses on the various contents of the general wish. The wish in the main

system is the rational faculty's desire for truth; the wish in the sub-system is the desire to believe that p; and over all stands the desire to avoid conflict. It may seem that another paradox is lurking in this, but as I have discussed earlier on, there are not multiplying wishes, but only one wish with multiplying effects. Pears suggests that perhaps the wish that lies at the heart of the sub-system should be specified in a more discriminating way. So, instead of naming it the wish to believe that p, it should be rephrased as "the wish that the main system should form the pleasure-giving belief," or as "the wish to eliminate the distressing belief from the main system". The main system faces a problem in that it looks onto a hostile environment, an environment that is constantly going to remind Emma of her sin. The wish in the sub-system, as part of the general wish to avoid conflict and distress, may have at its core a concern about the likely effect of that problem on the main system. In other words, the sub-system has as its environment the main system, and the elimination of anxiety and distress in the main system is what the wish in the sub-system has as its semi-altruistic aim. The main system, on the other hand, has as its environment the external world. So, when Emma's main system finds itself unable to dominate and change the environment and to undo the distressing event, the sub-system dominates the main system and eliminates the distressing belief from the main system. The paradox is avoided in that both systems have different environments, different inputs, aspirations and different problems, and because of all these differences the two systems receive and process information differently. This, in turn, implies that the two systems have different consciousnesses of their own (i.e. the functional meaning of "consciousness") which enable them to react to information. The sub-system has no direct access to the external world, so if it did have an internal consciousness, it would be one that was "buried alive." Pears notes that the unconscious often operates as if it has an internal consciousness of its own. (p.99). However, if

this consciousness if "buried alive", then because of its alienation from the external world and the resultant lack of evidence, neither the truth nor the falsity of this hypothesis can be determined.

Fingarette, like Pears, prefers the functional interpretation of the Freudian theory of consciousness and suggests that Freud himself favoured this view. Fingarette (1969, p.111) quotes the opening words of Freud's last paper entitled "Splitting of the ego in the process of defence":

> "I find myself for a moment in the interesting position of not knowing whether what I have to say should be regarded as something long familiar and obvious or as something entirely new and puzzling. But I am inclined to think the latter." (Freud, 1938, XXIII, p.275)

According to Fingarette, the "new and puzzling" concept of the defensive process is that it is not something that happens to the ego, but it is rather something the ego does, a motivated strategy. The productive cause in the defensive process is still the wish to avoid anxiety but, Fingarette maintains, Freud did not appreciate the importance of the rôle of the ego in the mode of operation of the defensive process until the "very last days of his life". Traditionally, the split was between Consciousness and the Unconscious, with the latter characterized by a "primitive" character. All elements that, therefore, split off from the ego take on a markedly "primitive" character. The problem for Freud with this interpretation is to explain how the split-off element still retained to a great extent the fundamental characteristics of the ego, or in Pears' terms, how the sub-system can be internally rational. To solve this problem, Fingarette postulates three principles on which the Freudian theory rests and then recasts Freud's theory in more functional terms. The three principles are:

1. The self-alienating characteristic of defence is reflected in an alteration of consciousness.

2. What is of great importance in defence is the "dynamic" aspect.

3. The alteration of consciousness in defence should be understood in terms of the "dynamics" of defence, rather than by reference to traditional terms like "knowledge" and "ignorance". (Fingarette, 1969, p.127).

Accordingly, Fingarette rephrases Freud's theory as follows:

> "The result of defence is to split off from the more rational system (i.e. the system which is defended) a nuclear, dynamic complex. This nuclear entity is a complex of motive, purpose, feeling, perception, and drive towards action." (p.129).

So, instead of conceiving the two systems as Consciousness and the Unconscious, he interprets the two systems as operating within the ego.(4) The two systems are the elaborated ego-structure and the split-off ego-nucleus. The reason for the split, according to Fingarette, is the incompatibility between the ego-nucleus and the current ego. The split is too great for the integrative capacities of the ego, or the self-synthesis of the person, that the ego gives up any attempt to integrate the ego-nucleus. In other words, the ego does not acknowledge the ego-nucleus which is experienced by the person as the "not-me"; the ego disavows that part of itself which cannot be integrated into the self-synthesized personal identity. According to Fingarette, the ego treats this unassimilable but ego-like system as "outside" rather than "inside". This is, of course, Fingarette's interpretation of Freud, an interpretation that gives support to his own view of self-deception as a form of disavowal (see Chapter 4). But perhaps Fingarette reads too much into Freud when the latter stated:

> "When this process (of defence) occurs for the first time there comes into being a nucleus and centre of crystallization for the formation of a psychical group divorced from the ego — a group

> around which everything which would imply an acceptance of the incompatible idea subsequently collects." (Freud, 1895 II, p.123 as quoted by Fingarette, 1969, p.132)

By recasting Freud's theory in terms of "disavowal", ego-rejection and "ego-synthesis" to support his own view of self-deception as "a refusal to spell out one's engagement in the world", Fingarette lays himself open to the same criticisms as before (see Chapter 4). It may be an interesting account of why self-deception takes place, but Fingarette fails to state just how it happens. Furthermore, I think that he is misrepresenting the Freudian theory when he states, "Freud .... eventually appreciated that his therapy had always been oriented primarily to self-acceptance .... rather than to 'knowledge' .... as curative. Avowal of one's engagements is the optimal goal of classical psychoanalysis." (p.142).

In the next chapter, I will show how Davidson's functional account of self-deception can be based on the exact same quote from Freud. Freud's theory can be interpreted also in terms of "knowledge" and "belief" and "reasons", an account that is primarily interested in how self-deception works rather than the why of self-deception.

* * * * *

In this chapter I have looked at a theory of self-deception that can account for "hard" cases of self-deception. The cases I have concentrated on are cases of self-deception in which the productive cause is a desire (i.e. a "hot" case of self-deception) and the permissive cause a failure of consciousness. Pears admits that Freud's theory "works perfectly as a theory about the permissive cause of the extreme cases of irrationality that would be impossible without it, and it can be extended to cases that would only

be difficult without it." (Pears, 1984, p.77). Freud's theory can, therefore, be applied to the majority of cases of self-deception, viz. "weak" cases where "a failure of consciousness is an important luxury rather than a necessity. In these cases the Freudian theory of systems still has considerable explanatory power, although, it must be admitted, less than in the extreme cases that could not occur without a failure of consciousness." (p.78).

I stressed the importance of distinguishing between the two meanings of "conscious"——the first, more traditional, meaning refers to the name of the system that controls our daily lives, and the second meaning to the inter-action that exists between this sytem, the main system, and a rebellious sub-system that is brought into being when the subject practises "hard" self-deception. The earlier theory built on the first meaning faces certain problems which the later theory built on the second meaning can answer. Fingarette maintains that Freud himself saw the shift in emphasis, but there is no direct evidence in Freud's writings of this. What I term the "later" Freudian theory is, in fact, the theory as interpreted by Pears, Fingarette and Clavell.

In terms of the early theory, Pears appeals to the notion of time to overcome the various paradoxes. Owing to the subject's avoidance of not-p because of the general wish to alleviate conflict, lack of supportive evidence for the belief that not-p either weakens the belief or causes it to fade gradually from consciousness. Once it is safely out of consciousness, the self-deceiver can rely on his wish for p to generate the belief that p in consciousness, a belief that can be held without conflict since the belief that not-p is safely screened by the censor. However, the problem that arises is that the second part of Freud's principle holds that when the rational belief that not-p and the cautionary belief are in the

unconscious, they have no power to prevent the formation of the irrational belief in consciousness. And yet, these beliefs do have an effect on consciousness and do guide it. Here Pears relies on the later theory which appeals to the internal rationality and consciousness of the rebellious sub-system. The sub-system is built around the wish for the irrational belief and everything within the sub-system, therefore, works towards actualizing the formation and maintenance of the irrational belief in the main system. The sub-system is the organizing centre which forms, plans and guides the whole strategy of self-deception. This is not necessarily the case in the initial stages during which the self-deceiver may intentionally plant the seeds for the plan of self-deception, trusting that the general wish for p will take over the nurture and care of these seeds. The self-deceiver cannot plan and envisage the actual growth of the mature plants, but relies on certain forces (i.e. "the discreet operation of his wish") to do the job. He, therefore, hands over responsibility at this stage so that when eventually, with time, the seeds of self-deception blossom into a full-grown state of self-deception, the agent has divorced himself from the process.

The sub-system, therefore, seems to be a highly organized and complex system, taking note of dangers, steering the main system away from not-p, screening the irrationality of the agent's belief from consciousness, etc.—pointing to a system that is both internally rational and conscious. But this threatens to be another fertile breeding ground for the epistemological paradox: if the sub-system is internally rational and conscious, how can it contain both the rational and cautionary beliefs and yet be built around the nucleus of the wish for the irrational belief? Pears suggests that the internal rationality of the sub-system is not threatened if the wish around which it is built is described as the wish for the well-

being of the main system, i.e. the wish for the main system to form the irrational belief that will give the agent satisfaction. With perfect rationality, the sub-system eliminates not only the rational and conflicting belief that not-p but also the cautionary belief from the main system, allowing the irrational belief to flourish unhindered and to give satisfaction.

It is precisely this later aspect of the Freudian theory on which Davidson builds his own Functional Theory —— the apparent rational relationship that exists between these two semi-autonomous sytems:

> "After analysing the underlying problem of explaining irrationality, I conclude that any satisfactory view must embrace some of Freud's most important theses, and when these theses are stated in a sufficiently broad way, they are free from conceptual confusion."
> (Davidson, 1982, p.290)

In the last chapter I shall look at how Davidson bases his theory on that of Freud and how the functional theory hopes to escape "conceptual confusion".

Notes

1. Cf Kipp (1980, p.309). He criticizes Freud unfairly by stating that the Freudian theory "seems to require that two mutually opaque, autonomously thinking and willing consciousnesses should exist within the soul of the self-deceiver, yet that these consciousnesses should also exist within a unified consciousness that grounds the self-deceiver's identity as a self. Without the first of these two conditions, properly deceptive concealment of belief and intention seems unthinkable, and without the second, properly self-deceptive reflexivity, within the relation between deceiver and deceived, seems equally unthinkable." Kipp's reference to the first condition refers to consciousness (indeed, consciousnesses) as a separate system, whereas the second "unified consciousness" refers to the relation that holds between the two systems.

2. Sartre stipulates that in self-deception the initial belief that not-p and the suppression of it must happen simultaneously and "not at two different moments." In Chapter 4 I discussed that this is not a necessary condition in successful lying to others. When A lies to B, B may first examine and mull over what A has told him before coming to believe it.

3. See Hamlyn (1971) who argues that the self-deceiver does know what he is doing, but he does not know it consciously. Self-deception is an intentional activity that involves a strategy. See also Audi who states that unconscious beliefs "tend to manifest themselves in consciousness and behaviour, and in essentially the same way as conscious beliefs" through, for example, slips of the tongue. (1982, p.137). Furthermore, he argues that it is as if the self-deceiver were two people because he operates at two levels: at the conscious level and at the unconscious level or "metalevel from which he manipulates his own consciousness or behaviour." (p.141).

4. It is important that "ego" is not equaled with the traditional meaning of "consciousness", since the ego can operate on different levels of consciousness. The ego is the "control centre" of the personality, either holding back or releasing the expression of basic drives, and attempts to reduce tensions by dealing successfully with the environment. Ego is part of a later Freud's dynamic structural model of personality, whereas consciousness is a level in the early Freud's static topographical model of the personality. The ego can operate on the unconscious level when it represses something which is unacceptable to the person.

CHAPTER 6

DAVIDSON'S FUNCTIONAL THEORY OF SYSTEMS

I ended the last chapter with Davidson's reference to Freud, and in this chapter I want to trace similarities in the two theories of systems, as well as the differences which, however, do not lead to the exclusion of one theory by the other, but rather to a complementing and unification of both theories into one coherent theory of self-deception. As Pears notes, "it is true that the functional theory is compatible with Freud's idea." (1984, p.88). As my starting point I am going to take the four fundamental Freudian doctrines on which Davidson bases his theory and shall trace his development of them in order to explain how irrationality, especially internal irrationality, is possible:

1. The mind contains a number of semi-independent structures, having mental attributes like thoughts and memories.
2. These systems can interact to cause further events in the mind, or outside it.
3. Some of the dispositions and events that characterize the various substructures in the mind are like physical forces when they interact with other substructures — a causal relation between parts in which reason does not play its usual normative and rationalizing rôle.
4. Unconscious mental states and events are like conscious beliefs, memories, desires, wishes and fears. (Davidson, 1982, p.290)

Before looking at how Davidson explains irrational action, it is necessary to establish what, for Davidson, constitutes standard reason explanations for rational behaviour. After all, irrationality is a breakaway from

rationality and should, therefore, be interpreted against a background of rationality. Davidson's pattern of reason explanations concentrates on the rationalist motif in his thought. To be rational is to have sufficient reasons for what one does, in thought and in action. According to Davidson, a reason is a mental cause which is a rational cause, at least in normal, intentional action. This view is first developed in Actions, Reasons, and Causes in which reasons consisted of motives. To be rational was to have a "primary reason" to rationalize one's actions. This "primary reason" was both a "pro-attitude" or desire and a belief about how to satisfy it. The primary reasons included some desire of the agent whether it was to touch an elbow or to pull out weeds or even to cheat one's son out of greed. (Davidson, 1963, p.7)

> "In the light of a primary reason, an action is revealed as coherent with certain traits, long - or short-termed, characteristic or not, of the agent, and the agent is shown in his rôle of Rational Animal .... Central to the relation between a reason and an action it explains is the idea that the agent performed the action because he had the reason." (p. 8,9).

In How is Weakness of the Will Possible? Davidson regards reasons as premises, and these reasons must be good enough to ward off charges of irrationality. This rationalist view of reasons as premises, and of reason-guided activity as the making of valid moves from all the available premises, develops the earlier view of a "primary reason". In Paradoxes of Irrationality Davidson further develops the pattern of reason explanations. He looks at both the causal rôle as well as the logical relation of reasons to actions. He quotes Hume in order to illustrate the pattern of reason explanations.

> "'Ask a man why he uses exercise: he will answer, because he desires to keep his health. If you then enquire why he desires health, he will readily reply, because sickness is painful.'"
> (Davidson, 1982, p.293)

If we are to explain the person's taking exercise, the explanation will have to include at least two factors: (1) the "pro-attitude" of the agent, a desire for health and, (2) a belief that in acting in a certain way he will realize his desire and be satisfied, a belief that exercise will make him healthy and that there is, therefore, something desirable about taking exercise. In this way, his taking exercise is explained in terms of reasons, a belief-desire pain. However, this belief-desire pain is related to the action in two ways to yield an explanation. Firstly, beliefs and desires have a content, a content about the desirability of health and the way in which to fulfil this desire by acting in a certain way, and these contents imply that there is something desirable about taking exercise. There is, therefore, a <u>logical</u> connection between the contents of the desire and belief and the action of taking exercise. Secondly, the reasons the agent has for acting in a certain way must be the reasons for that action, i.e. he exercises because he wants to be healthy and because he believes that exercise promotes health. In other words, the reasons, i.e. the desire and belief, must play a <u>causal</u> rôle in bringing about the action.

> " .... there is no inherent conflict between reason explanations and causal explanations. Since beliefs and desires are causes of the actions for which they are reasons, reason explanations include an essential causal element." (p.293).

Davidson stresses the causal aspect of intentional, rational action by referring to our wishes, hopes, desires, beliefs, thoughts, etc. which depend on simple inference from other beliefs and attitudes. We believe that exercise promotes health on the basis of induction from hearsay or reading or personal experience. The action of exercising is intentional and, regardless whether this intention to exercise is executed or not, the intention itself is caused by a desire to be healthy and a belief that by

exercising we will be healthy. There is, therefore, both a logical (or rational) and a causal aspect of the intention itself. It is this logical connection between the contents of various pro-attitudes and beliefs and what they cause which forms the basis of Davidson's functional theory. Thus, according to Davidson, all intentional actions have a rational element at the core. For the action to be intentional, the belief-desire pair must, of course, cause the action in the right way. In Chapter 1 I noted that the belief-desire pair can cause the action in various ways, but for A to deceive B intentionally, the belief and desire cause action x in the right way. For example, the man with a rope on the mountain has the desire to kill his companion and the belief that by dropping the rope the friend will be killed. He may become so agitated by his wicked thoughts that he drops the rope. In this case there is a belief-desire pair which causes his friend to be killed, but he did not intentionally kill his friend because his belief and desire did not cause the dropping of the rope in the right way. It is only in the case of his intentionally dropping the rope, knowing the other will be killed (and is killed) that the killing is an intentional action; in other words, that the belief-desire pair caused the action in the right way. There is no inherent conflict between reason explanations and causal explanations (at least, in rational behaviour). Since beliefs and desires are causes of the actions for which they are reasons, reason explanations include an essential causal element. In other words, in standard reason explanations not only do the propositional contents of various beliefs and desires bear appropriate logical relations to one another and to the contents of the belief, attitude or intention they help to explain, but the actual states of belief and desire also cause the explained state or event in the right way.

If this is the basis of standard reason explanations for rational action, what then is irrational action? In order to give a comprehensive account

of irrationality, it is necessary to appeal to the principles of rationality and intentional rational action. As Davidson points out in Paradoxes of Irrationality (p.303), the central paradox of irrationality is that it needs to be placed in a structure that allows for the inconsistent and unintelligible which is the central feature of irrationality. However, if too much incoherence is allowed into the system, there is the danger of removing the entire theory from the background of rationality which is needed as a yardstick against which we can measure the degree of the perversion of rationality. On the other hand, if we explain a theory of irrationality too well in a too structured system, we may turn it into a disguised or merely alternate form of rationality. This is the first paradox with which Davidson grapples: how is it possible to reconcile an explanation that shows an action, belief or emotion to be irrational with the element of rationality inherent in the description and explanation of all such phenomena. "The difficulty in explaining irrationality is in finding a mechanism that can be accepted as appropriate to mental events and yet does not rationalize what is to be explained." (Davidson, as quoted by Clavell, 1986, p.503). In other words, there is a conflict between the standard way of explaining intentional action and the idea that such an action can be irrational. The view that no intentional action can be internally irrational stands at the one extreme in the scale of possible views and is labelled the "Plato Principle" by Davidson.(1)

> "Someone who knowingly and intentionally acts contrary to his own principle; how can we explain that? The explanation must, it is evident, contain some feature that goes beyond the Plato Principle; otherwise the action is perfectly rational. On the other hand, the explanation must retain the core of the Plato Principle, otherwise the action is not intentional."
> (Davidson, 1982, p.297)

The psychological state or event of self-deception entails what is loosely

called a propositional attitude which points to the relevance of a reason explanation and thus to the element of rationality in self-deception. However, self-deception is irrational and the element of rationality cannot prevent its being at the same time less than rational. In other words, to have a reason for believing an irrational belief (e.g. because one desires that p were the case) entails an appeal to rationality or a deviation from it.

This brings Davidson to the second problem in that if a cause is described in non-mental terms, it loses touch with the element of rationality. "Events conceived solely in terms of their physical and physiological properties cannot be judged as reasons." (p.299). These neutral forces have no mental status as beliefs or attitudes, but are external to the mind, according to Davidson, and therefore cannot be part of the rational or irrational. They are part of the non-rational. The conclusion is thus that the description of irrationality must entail a mental description of a mental cause, thereby making it a candidate for being a reason. By appealing to both rationality and causation, Davidson hopes to reconcile the two tendencies in Freud's theory. "On the one hand he (Freud) wanted to extend the range of phenomena subject to reason explanations, and on the other, to treat these same phenomena as forces and states are treated in the natural sciences. But in the natural sciences, reasons and propositional attitudes are out of place, and blind causality rules." (p.292). Davidson hopes to bridge this gap by looking at the causal element of reason explanations. Hence his search for a "mechanism that can be accepted as appropriate to mental events."

However, I have not yet answered the question posed earlier on, namely what then is irrational action? and how is internal irrationality possible ?

The answer is found in Paradoxes of Irrationality in which Davidson discusses at length the idea that irrationality always incorporates a mental state which causes another mental state but is not a reason for it. The causal link remains in that one mental state causes another, but the logical relation is distorted, for in "hot" self-deception a wish is not a justified cause for a belief.

Most of Davidson's theory of irrationality is applied to cases of akrasia. I shall, therefore, first examine his reason explanation for akrasia, extract general principles and then apply these to cases of self-deception. Let me start with an example of rational behaviour, based on Davidson's example of the man in the park (1982, p.292ff), develop it into a case of irrational action, and then see where the mental breakdown in rationality occurs which gives rise to irrationality. Tim has discovered that his friend Eric has stolen some money from him. Very disappointed, Tim writes Eric a letter expressing his dismay at his friend's disloyal deed, and deposits this letter in Eric's post box. On his way back home, he feels that the sum was a negligible amount, and not worth the loss of a friendship. He returns to Eric's house and retrieves the still unopened letter from the box. Here everything Tim does is done for a reason, which makes the corresponding action reasonable. In each case the reasons for the actions tell us the intention with which Tim acted, and thereby give us a reason explanation of the actions. As I noted before, there are two factors involved in reason explanations. First of all, there is a value, wish or attitude of the agent — Tim wants to communicate his disappointment to Eric. Secondly, there is a belief that by acting in the way to be explained, he can promote the relevant value or satisfy the relevant desire — Tim believes that by delivering the letter he will make known his feelings to Eric. Furthermore, these two factors, the action on the one hand, and

the belief-desire pair on the other, are related to each other in two different ways in order to yield a standard reason explanation. Firstly, there must be a logical relation. Beliefs and desires have a content, and the content must be such that it implies that the action will satisfy it. Tim has a desire to save the friendship; he believes that by removing the recriminating letter he will save the friendship, and he concludes that there is something desirable in returning to Eric's, which is his reason for going back. Secondly, Tim went back to retrieve the letter because he wanted to save the friendship. In other words, the reasons play a causal rôle in the occurence of the action.(2)

How can Tim's action be interpreted as being irrational? Before returning to retrieve the letter he has deliberated on two issues: he has a motive (the desire to remain friends) for taking away the letter, but he also has a motive for leaving the letter there (the desire to communicate his disappointment to Eric). If in his own judgment, the former consideration outweighs the latter, he will be acting in accordance with his final value-judgment (i.e. to save the friendship). If, however, in his own judgment the latter consideration outweighs the former, and yet he acts on the former, he will be acting against his own better judgment. In other words, the akratic act is irrational. How has it become irrational?

And this brings me to an extremely important distinction made by Davidson. Tim's action may appear akratic to an observer, i.e. external irrationality, but there is not necessarily an inconsistency by Tim's own standards, in other words, there is not necessarily internal irrationality. Before developing Tim's case into one which reflects internal irrationality, I want to repeat the distinction made at the beginning of Chapter 3. External irrationality is not conceptually problematic, whereas internal irrationality

does seem to involve a paradox.

> "Much that is called irrational does not make for paradox. Many might hold that it is irrational, given the dangers, discomforts, and meagre rewards to be expected on success, for any person to attempt to climb Mt. Everest without oxygen (or even with it). But there is no puzzle in explaining the attempt if it is undertaken by someone who has assembled all the facts he can, given full consideration to all his desires, ambitions and attitudes, and has acted in the light of his knowledge and values. Perhaps it is in some sense irrational to believe in astrology, flying saucers, or witches, but such beliefs may have standard explanations if they are based on what their holders take to be evidence. It is sensible to try to square the circle if you don't know it can't be done. The sort of irrationality that makes conceptual trouble is not the failure of someone else to believe or feel or do what we deem reasonable, but rather the failure, within a single person, of coherence or consistency in the pattern of beliefs, attitudes, emotions, intentions and actions." (p.290).

Tim's either negative or positive evaluations are what Davidson terms conditional or prima facie judgments. This judgment is, of course, not necessarily conclusive. Davidson in How is Weakness of the Will Possible? distinguishes between these prima facie or conditional evaluative judgments (e.g. "With all the evidence carefully considered, I ought to do x") and unconditional judgments (e.g. "I will do y"). The "all things considered" or conditional judgment does not come into conflict with unconditional judgments, because to judge an action desirable is not yet to judge it more desirable than any alternative. The move from "It would be best to do x" to "I shall do y" does, therefore, not necessarily involve inconsistency.

> "Our 'best' judgments .... could naturally be taken to be those conditioned on all the considerations deemed relevant by the agent; but an action is geared to unconditional judgments. Since there is no principle or psychological law that says we must trim our unconditional judgments of what is best to

> our best judgments, someone can judge, and act, contrary to his own best judgment." (Davidson, 1985 b, p.201)

In other words, the akrates who judges "It would be better not to do y than to do y; but y is desirable and I shall do y" is not yet guilty of inconsistency in judgment. To reduce the akrates's supposed irrationality to inconsistency requires, according to Davidson, an extra higher-order premise, the judgment that "I ought to act on my own best judgment, what I judge best or obligatory all things considered."

Therefore, a person who is aware of the fact that he has good reasons both for and against an action is not necessarily entertaining a contradiction. Most actions we perform, or consider performing, have something to be said both for them and something against. We speak of conflict only when the pros and cons are so closely weighted and balanced, as to make a choice difficult. If I have accepted an invitation (and wish to keep my promise) for a dinner party on Friday evening and if I want and am expected to attend a friend's funeral on Friday evening, simple logical deduction will tell me that I'll have to be at two different places at the same time, but logic cannot tell me which to do. Since logic cannot make the choice for me, the question may arise then in what respect either action would be irrational. Even if I add the condition that all things considered I ought to go to the dinner party, yet attend the funeral instead, it is still unclear whether the irrationality is evident. Inconsistency does not arise if I have only the two judgments: the conditional judgment that in the light of all my evidence I ought to attend the dinner party, and the unconditional judgment that I will attend the funeral. Pure internal inconsistency enters only if I hold, what Davidson calls, my second-order principle: that I ought to act on my own best judgment, everything considered. It is this description that makes an akratic act irrational and it is only when we can describe an

action in just this way that it becomes problematic. When I discussed Davidson's view of akrasia in Chapter 2, I noted the fairly extensive quote in which Davidson argues for the distinction between internal and external irrationality. For the action to be internally irrational the agent must go against his own second-order principle. If, on the other hand, he does not hold this principle, there is nothing necessarily paradoxical about his action which to us, as onlookers, may, however, be judged to be irrational, i.e. externally irrational. To explain the agent's action we need say only that his desire to do x (that which he considers best) was not as strong as his desire to do y (the externally irrational action).(3)

That returns us to the initial problem which is now phrased differently. Instead of asking, "How do we explain someone who knowingly and intentionally acts contrary to his own better judgment, all things considered, and who is capable of avoiding irrational action?" the question now is, "How do we explain someone who knowingly and intentionally acts contrary to his own second-order principle?" When Tim returns to retrieve the letter he has a reason: he wants to save the friendship and he believes that by destroying the letter he will do so. But in doing so, he ignores his second-order principle that he ought to act on what he thinks best, all things considered. The motive for ignoring his principle (that he ought to act on his better judgment that Eric should know that Tim is aware of his disloyalty and that he should thus leave the letter) is the very strong desire to retrieve the letter and so save the friendship. And this is the point at which irrationality enters. The desire to save the friendship has entered into the rational decision. It was a rational decision, all things considered, not to return to Eric's, and given this principle, Tim ought to have acted on such a conclusion. But he doesn't. Irrationality entered when his desire

to return made him override his principle. Although his motive for ignoring his principle is a reason (a motive) for ignoring the principle, it is not a rational reason against the principle itself. And so, when it enters in this way, it is irrelevant as a reason, to the principle and to the action.

Davidson's interpretation of this kind of irrationality depends on the distinction between a reason for having, or acting on, a principle and a reason for the principle itself. I shall use an example which is analogous to Davidson's example of the man desiring a well-turned calf. (p.298). A young man very much wishes that he were handsome and this leads him to believe that he is. His reason for having this belief is that it gives him pleasure. However, if he holds this belief <u>purely</u> because he wants to believe it, then his holding the belief is irrational. Wishing to have a certain belief does in no way contribute to the truth of that belief. It is for this reason that a wish is not a reliable cause for a belief (except, as I have noted earlier in Chapter 2, for the belief that one has that wish). The wish to have a belief gives no rational support for the truth of the belief itself, but what it does make rational is that this proposition should be true: He believes that he is handsome. This, however, does not give rational support for his belief: I am handsome. This is a case of irrationality. His desire, therefore, is a reason for having the belief (it gives him pleasure), but it is not a reason for the belief itself (it is not a justified cause). So, his desire to be handsome causes his belief, but the desire is not a reason to believe that it is so; there is no logical relation between his desire and his belief. The irrationality does not lie in the belief itself, but in the fact that it is caused by a wish. This is an important point and I shall discuss later how Davidson's example of irrationality is actually ambiguous.

As noted before, in standard reason explanations the propositional contents of various beliefs and desires have a logical relation to one another, and the belief-desire pair cause the action. However, in the above example, a desire caused a belief, but the judgment that a certain state of affairs would be desirable is no rational reason to believe that it exists. "In the case of irrationality, the causal relation remains, while the logical relation is missing or distorted." In other words, "there is a mental cause that is not a reason for what it causes." (p.298). And so, Davidson bridges the gap that posed a problem for Freud ——the causal relation that holds between mental events—— as well as the problem posed by the Plato Principle ——the explanation of an irrational belief still retains an element of rationality in that the irrational belief is explained by a mental cause that is not a reason for it.

I noted earlier on that since the bulk of Davidson's writing on irrationality is addressed to the problem of akrasia, I would briefly examine this phenomenon and then extend the principles of the theory to the concept of self-deception. The similarity is that both the weak-willed agent and the self-deceiver operate in a context of mental conflict. The akrates acts intentionally on one or a few favoured reasons, and not on all the reasons that are recognized as being relevant to his situation. He has reasons both for and against a certain cause of action; deliberates on the reasons for his action; but then discards them in favour of "better" reasons for another course of action. This is the cognitive aspect of weakness of will and which Davidson refers to as "weakness of warrant." The agent, with all the evidence available to him, will judge a certain hypothesis more probable than not, and yet he goes against this hypothesis. The agent, like the self-deceiver, has to decide between two mutually exclusive hypotheses: the hypothesis supported by relevant evidence and the negation of the hypothesis.

The analogous weakness of the warrant enjoys the same logical structure (or, as Davidson points out, illogical structure) as weakness of will. The cognitive error in weakness of the warrant is an irrational belief which goes against the available evidence, whereas weakness of will involves an irrational intention or intentional action which goes against the values the agent holds. "Weak" self-deception then, according to Davidson's argument, includes weakness of the warrant in that an irrational belief is held, even though the self-deceiver has better reasons for accepting its negation. "Weak" self-deception, therefore, is analogous to akrasia. However, the "hard" self-deceiver is guilty of irrationality on at least one further count: the fact that his favoured belief is caused solely by his desire for that belief; there is no evidence to support his favoured belief. In other words, the desire is a mental cause which is not a reason for the favoured belief. Whereas in akrasia there is no "psychological law that says we must trim our unconditional judgments of what is best to our best judgments" (Davidson, 1985 b, p.201), there is, for us rational people a psychological law which encourages us to trim our beliefs to the available evidence. There is nothing intrinsically irrational about the akrates who intentionally performs an action purely because he so desires, but there is something intrinsically irrational about the agent who intentionally forms and holds a belief purely because he so desires, even though he knows, as a rational person, that rational beliefs should be based on the available evidence. And this is what makes self-deception more problematic than akrasia.

As I noted in Chapter 4, "weak" self-deception, according to Davidson is not necessarily internally irrational on this first level since, like in weakness of the warrant, there are reasons both for and against a certain proposition. There are better reasons for accepting the negation of the

proposition rather than the proposition he does accept. The negation of the proposition rather than the proposition the self-deceiver accepts is, therefore, more likely to be true, but the self-deceiver bases his acceptance of the proposition on what he takes to be only part of the relevant evidence. To reduce "weak" self-deception to inner irrationality would require that the self-deceiver also holds a second-order principle, the "normative principle against which such a person has sinned .... called the requirement of total evidence for inductive reasoning: when we are deciding among a set of mutually exclusive hypotheses, this requirement enjoins us to give credence to the hypothesis most highly supported by all available relevant evidence." (Davidson, 1985, p.140). However, like the second-order principle for akrasia, this principle for inductive reasoning is not always operational. Davidson notes (p.141) that someone who acts or reasons in accord with these principles does not do so at all times, otherwise internal irrationality would not be possible. Nor does the person who acts or reasons in accord with these principles do so seldomly or never: to accept the principle is to act or reason in accord with it most of the time. Davidson adds the condition that the acceptance of the principle consists mainly in that person's "pattern of thoughts" being in accord with the principle. This does not mean, however, that the person who accepts the principle of the requirement of total evidence must be constantly aware of it.

However, in this chapter I am concerned mainly with "hard" self-deception. Davidson has shown that many of the things we might mean in calling a thought process or an action irrational do not involve paradox. "Hard" self-deception is a case of incoherence within a single person in the pattern of his beliefs and desires, and it is this inner irrationality that gives rise to conceptual difficulty. The paradox is brought about by the

fact that desires, beliefs and the actions they explain, are distinguished and identified partly by their logical relations with one another, that is, rational relations. These logical relations can also be described in terms of certain second-order principles such as "Believe that proposition for which there is the greatest amount of evidence", and "Perform that action you think is best, all things considered". The problem is that akrasia and self-deception seem to be cases precisely of not believing or acting in accord with the over-arching principles of rationality or coherence. So the "hard" self-deceiver is guilty on two counts: firstly, he clings to a belief for which there is no evidence (as opposed to the "weak" self-deceiver who bases his belief on only part of the evidence) and the belief is caused purely by a desire that acts as a mental cause but is not a reason for the favoured desire; and secondly, he goes against his second-order principle which, as a rational person, he must hold. The self-deceiver has a "reason" for this weakness of the warrant, but it is a "reason" which he himself has brought about. Self-deception is self-induced. The self-deceiver's "reason" is his desire to believe that p and is, therefore, not a rational reason, but it is a cause of the belief that p.

In the example of Emma, the mental cause for her belief that p (i.e. the belief that she does not have an illegitimate child, that she is virtuous, etc.) is her desire for p. This desire acts as a "reason" for forming the belief that p, but it is not a rational reason for forming the belief that p against the second-order principle of "Believe that proposition for which there is the greatest amount of evidence", nor as a rational reason or "justified" cause for bringing about the belief that p. So the desire, or mental cause, has to operate in part as an irrational reason. Baier criticizes Davidson on the ground that in his theory a mental cause which operates in part irrationally is not to be regarded as a reason in

respect of that part of its operation, but only as a cause that is not a reason. Pears (1985) defends Davidson's view by replying to Baier's criticism that it is true that the mental cause is irrational or "surd" at that point in its operation, but this does not mean that it is, therefore, not a reason. Pears points out that Davidson means that the mental cause does not operate as a rational reason and not that it does not operate as a reason at all. Davidson himself does not make this clear distinction and so his dictum that in irrationality "a mental cause does not operate as a reason" may lead to confusion. He does, however, note the "ambiguity of the phrase 'reason for believing'" in Paradoxes, but it is only later in Deception and Division that he elaborates on the different application of "reason". "Charles has a reason to believe that p" is ambiguous. "A reason of the first sort is evaluative: it provides a motive for acting in such a way as to promote having a belief. A reason of the second kind is cognitive: it consists in evidence one has for the truth of a proposition." (1985, p.143). In other words, we could see the evaluative reason or motive as a causal (albeit irrational) reason, and the second cognitive reason as a logical (and, therefore, rational) reason. Davidson's earlier explanation, however, gives the impression that reasons are necessarily rational for he says more than once that, when a mental cause is operating irrationally, it is operating as a cause that is not a reason. What he means is that it is then operating not as a rational reason, but as an irrational reason.

Pears draws attention to two easily confused, but in fact two quite different, distinctions. First of all, Davidson should distinguish between a mental cause that operates non-rationally and one that operates as a reason, i.e. rationally. An example of a mental cause operating non-rationally is Davidson's own example of a person humming a tune and it reminds him of a name. The point that is being made here is that non-

rational mental causation is not necessarily irrational. The humming of the tune is a simple case of association. Here the humming is a mental cause for the recollection, but not a reason for it, at least not in the sense of "my reason for it", or a "good" reason for it. Nevertheless, this is not a case of irrationality.

Davidson offers a second, more difficult, example to substantiate his view that being in a mental state which has a mental cause which is not a reason (or rational reason) is not a sufficient condition for irrationality. An agent sets out to change his own character by taking a certain cause of action. There are several ways in which this may come about. Davidson focuses on a case in which the agent's "reason" for changing his present character is based on a value that he does not yet accept. In this case, the value that produces the change is extrinsic to his present character and so, although it operates as a cause in his development, it does not operate as a rational reason for it. And yet this type of action, a type of self-criticism or self-evaluation and reform, seems to be the very essence of rationality, an action which is held to be the basis for rational growth. I shall illustrate Davidson's point with an example: I desire to change the gluttonous aspect of my character, viz. my desire for chocolate cake. My desire to change must be operating from without the content of the value that is to undergo change. From the point of view of the changed value there is no reason for the change, since it comes from an independent source. So when the changed value is brought about, the desire to change, although it caused the changed value, cannot be a rational reason for what it caused. Davidson is impressed by the fact that the value operates from outside the part of the agent's character that he wants to alter so that, in a certain sense, the change originates in a different system to that which is being changed. This leads to the question that will introduce

the second part of Davidson's theory: must this different system be another part of his character? With this last example Davidson should, in order to avoid misinterpretation, have made the distinction between mental causes that operate as rational reasons (as in all rational behaviour) and mental causes that operate as irrational reasons. It is this very operation of a mental cause as an irrational reason that leads to the central question in Paradoxes of Irrationality: "How is internal irrationality possible?"

Davidson's answer is that human agents are often divided against themselves. He wants reasons (non-rational, rational and irrational) to be mental causes, and wants to characterize irrationality as having a mental, so a reason-like, cause of one's action which, nevertheless, fails to have the right "logical relation" to its effect to be a rational reason. In this way, Davidson's theory retains the core of rationality — the Plato Principle — which is imperative if a theory of irrationality is to make any sense at all. Davidson looks to social interaction for a model of that sort of mental cause, and then applies the model to a single agent. Davidson's gardener (1982, p.300) wants another person to enter his garden, so he grows a beautiful flower there. The other person craves a look at the flower and consequently enters the garden. The gardener's desire caused the other's craving and action, but the gardener's desire was not a rational reason for the other's craving, nor one on which the other person acted. Here the act of entering the garden was done intentionally and not against the will of that person. The mental cause, namely the designing will of the flower-grower, brought about its effect, namely the other person's entering the garden, but it was not a rational reason for the other person's action.

Davidson suggests 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 that interact"(p.300). Thus, a single person is divided into two systems in order to explain the formation of an irrational belief that he is competent to detect and avoid. The two systems, main system and sub-system, must be to some degree independent if we are to understand how they harbour inconsistencies, and how they interact on a causal level. The explanation must appeal to the supposition of two semi-autonomous departments in the mind: "one that finds a certain course of action to be, all things considered, best, and another that prompts another course of action. On each side, the side of sober judgment and the side of incontinent intent and action, there is a supporting structure of reasons, of interlocking beliefs, expectations, assumptions, attitudes and desires." (p.300). Davidson's idea of mental compartmentalization is one of overlapping territories, but with a degree of independence, for how else then would reason be defeated? He emphasizes that his idea of mental compartmentalization should not be confused with phrases like "partition of the mind" or "segment of the mind" which may erroneously be taken to suggest that what beyongs to one part of the mind cannot belong to another. "Quasi-independent sub-systems" are postulated within one person at the time of irrationality. The person's total system of beliefs and desires, second-order principles (which dictate what is to be done when desires conflict) splits into a rebellious sub-system consisting of the desire and beliefs relevant to its satisfaction. "The sub-system is built around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate centre of agency within the whole person, it is, from its own point of view, entirely rational." (Pears, 1984, p.87)

The question which necessarily arises is, "When does such fragmentation

occur?" According to Davidson, "the breakdown of reason-relations defines the boundary of a sub-division." (Davidson, 1982, p.304). This is because in a case of irrationality, "there is a mental cause that is not a reason for what it causes. So, in wishful thinking, a desire causes a belief. But the judgment, that a state of affairs is, or would be desirable, is not a (rational) reason to believe it exists." (p.298). The parts are defined in terms of function, in terms of the operative concepts of reason and cause. Davidson's criterion for the boundary between main system and sub-system reflects the interaction between the attitudes, desires and beliefs of a person rather than the Freudian criterion which involves consciousness of them. If, for example, an agent's belief that it would be irrational to indulge in a particular piece of wishful thinking fails to intervene in the main system and fails to prevent him from indulging in it, that belief is assigned to a sub-system. This is what Pears terms the cautionary belief. When the line between the two systems is drawn in this way, the result is a functional theory, because it is the actual functioning of the cautionary belief that decides on which side of the line it should be placed. The belief is cautionary and its proper function is, therefore, to intervene and stop the irrationality, but what it actually does, according to Pears, is "to sit on the sideline and let it happen." (Pears, 1984, p.69).

Reasons, as developed in How is Weakness of Will Possible?, are assimilated to premises in valid arguments, all contained within one arguer. The main system is seen as an autonomous reasoner, yet the sub-system with its rebellious desire produces its own practical syllogism. Both systems, therefore, retain their essential core of rationality. If both systems display internal rationality, the question arises of how irrationality comes about. Davidson uses the model of inter-personal relations in his

account of reasons for actions to illustrate the causal relations between sub-system and main system. He shows us how a mental cause of a mental effect need not be a rational cause, when the mental cause is in one person and the mental effect in another (the example of the flower-grower). Cases of social interaction seen as causal links between autonomous structures, are likened to the outcome of the conflict between a motivating desire and the principle to do what is judged best to do or to believe that for which there is most evidence. If someone's powerful will overcomes another's principled resistance to it, the first person's will may cause the other's surrender, so mental causes of mental effects may fail to be rational reasons at the interpersonal level. Davidson suggests that we use this as a model for understanding the akrates and self-deceiver. Here too, he suggests, are mental causes operating across the boundaries of semi-autonomous structures, mental causes which fail to be rational reasons. In other words, the causal relation remains but the logical relation is distorted. The division has nothing to do with consciousness as in the early Freudian theory of the divided mind. An element, according to the functional theory, is assigned to a sub-system whenever it fails to interact rationally with any element in the main system. The division depends on the function of the belief. The main system consists of all the desires and beliefs in the subject's psyche that interact with one another in a rational way to produce further desires and beliefs, and eventually speech and action. If one of the desires or beliefs fails to interact in a rational way with any element in the main system, it is banished to a sub-system. This sub-system, however, does not appeal to lack of consciousness.

Davidson's criterion seems to suggest that the desire or belief is assigned to a sub-system whenever it is guilty of non-intervention in the main

system; when it fails to intervene in the main system in some rational way in which it ought to intervene. Pears (1982, p.94ff) makes an important distinction. He refers to the failure to intervene rationally as the negative version of the criterion, but then postulates his own positive version: "It might mean that a desire or belief is assigned to a sub-system when it does interact with some element in the main system, but in an irrational way." (p.94). There is, therefore, a distinction between these two faults: the difference between irrational intervention and failure to intervene rationally. If we use the positive version, a wish that causes an irrational belief will be assigned to a sub-system simply because of its objective irrationality or irrational causation. If this version of the criterion is used, we need not even inquire whether the person is competent to detect and avoid the irrationality, or whether he possesses the cautionary belief which, nevertheless, fails to intervene and stop the formation of the irrational belief. The irrational efficacy of the wish, whether the person is aware of it or not, is sufficient to assign the wish to the sub-system. When this version of the criterion is used, Pears points out, a sub-system will be needed for any kind of irrationality.

However, if the negative version applies, we are dealing with irrationalities that the person can detect and avoid, the kind of cases of irrationality that interest Pears and Davidson. For in these cases, a sub-system will be needed only to house an element that belongs to the agent's psyche, but fails to produce the effect that it ought to produce, namely the inhibition of the irrational belief. In my example of the young man who wishes he were handsome, which is an adapted example of Davidson's man with the well-turned calf, it is not clear which version has been used. In this example the irrationality does not lie in the relation between

belief and evidence, but in the fact that his wish caused the belief—— and it is obvious that wishful thinking is irrational. So, it is not clear whether the sub-system in this example was brought about because of the objective irrational efficacy of the wish (the positive version) or whether the sub-system was brought about because the cautionary belief failed to intervene rationally (the negative version).

Pears states (p.98) that Davidson, in discussion with him, had told him that the negative version of the criterion is the one that should be developed, the version that applies to cases of irrationality that the agent is able to detect and avoid. Not only that, but Davidson is concerned with internal irrationality, which supports the negative interpretation. Pears notes several reasons for preferring the negative version. Firstly, the positive version postulates a sub-system in any case of irrationality. However, if the psyche of the agent cannot see the conflict between two incompatible elements, then there is no reason to suppose that they have to be kept apart, to avoid conflict, by being designated to two different systems in his psyche. It is only when the psyche is competent and able to detect the irrationality, that the sub-system comes into being (the negative version) to avoid conflict. Secondly, in the positive version the wish that causes a belief is banished to the sub-system because of this fault, its objective irrational efficacy. The wish cannot be placed in the overlap between the two systems (Davidson's compartmentalization is that of "overlapping territories") because the purpose of the positive application is to keep the wish and the belief apart for the same general reasons for which two contradictory beliefs have to be kept apart, namely because their relationship is objectively irrational. However, Davidson stresses that the wish and the belief <u>do</u> interact because the wish causes the belief and gives satisfaction in the

main system, but only across a line that marks the failure of their rational interaction. Davidson, therefore, places the wish, like Freud, in the overlapping section of the two semi-autonomous systems, one causally operative wish that belongs to both systems. Pears' other reasons for preferring the negative version include the productive inventiveness of creative thinking which often relies on non-rational association of ideas. Surely when the origination of a belief is irrational there need not always be a sub-system. Creative thinking would be severely inhibited if only linear rationality were always required.

But how then is Emma's case interpreted by the functional theory? Emma has good reason to believe that she has an illegitimate child (the personal experience, the photographs, etc.). However, this thought is painful and, therefore, to be avoided by her. The awareness of her unwelcome belief galvanizes her plan of self-deception, she is motivated to instill in herself a belief that she is virtuous, and she is motivated in the process of self-deception to satisfy that belief. However——and this is where the difficulty in explaining self-deception surfaces——once she has successfully reached the state of self-deception she must be motivated constantly to maintain this state which is continually threatened by evidence or even her memory. Paradoxically, her motivation is, therefore, based on the fact that the evidence points to her having had an illegitimate child. It seems as though, by a complicated circuitous route, we are right back to the epistemological paradox, i.e. the simultaneous entertainment of incompatible beliefs. But a successful state of self-deception is possible for there is a point in the sequence that led to Emma's state of self-deception in which there was a mental cause that was not a rational reason for the mental state it caused. Obviously Emma's two conflicting beliefs are kept apart, in two different systems.

Davidson, like Pears, holds that it is quite possible to hold incompatible beliefs but that it is not possible to conjoin them. Of course, the vast majority of Emma's beliefs are shared by both systems (in the large overlapping territory) but the contradictory beliefs cannot belong to the same system, for that would destroy the irrational belief. But, I have not yet noted where in the sequence that led to her successful state of self-deception there was an irrational step. The self-deceptive state consists of Emma's holding two conflicting beliefs and the step that made this possible is, therefore, the irrational point in the sequence: Emma's drawing of the boundary that keeps the inconsistent beliefs apart. The negative version of the criterion for drawing the boundary is used in this example, for the irrationality exists in the relation between belief and evidence and not merely in the fact that the belief is caused by a wish. The cautionary belief that she is being irrational, the second-order principle that she should hold that belief for which there is most evidence, the initial belief that she has had an illegitimate child and which motivates its own negation are all "walled off" from the rest of Emma's mind. What causes the sub-system to be walled off is her desire to avoid accepting what her second-order principle counsels. However, this desire is not a rational reason for neglecting her principle. "Nothing can be viewed as a good reason for failing to reason according to one's best standards of rationality." (Davidson, 1985, p.148)

I trust that it is evident at this stage how Davidson's theory overcomes the four main paradoxes postulated by Pears. First of all, it is possible for the self-deceiver to believe that p and believe that not-p at the same time, since the two beliefs are relegated to two different systems that fail to interact rationally. The second paradox, in which the intention must be screened, is one which Davidson addresses without

invoking the notion of consciousness. It is paradoxical that the state of self-deception is constantly motivated by that about which the agent deceives himself. In other words, the fact that the evidence points to the unwelcome belief motivates the self-deceiver to perpetuate his state of self-deception. This is made possible by the agent's drawing of the boundary which brings about a breakdown in reason relations and, therefore, allows the irrationality to remain unchecked in the main system. The drawing of the boundary is caused by the desire to avoid following the rational path that points to the acceptance of the unwelcome belief and this is in itself an irrational act. The Sartrean criticism of an infinite regress does, therefore, not apply since the boundary signifies a collapse of rational relations —— a boundary that allows all sorts of irrationalities to operate unhindered. The third paradox addresses the problem that continuing motivation is needed if the self-deception is to be successful. In order for the state of self-deception to be maintained it needs to be motivated by the desire to avoid accepting the second-order principle, but this is walled off from the rest of the self-deceiver's mind. So, it seems as though what is needed for a successful state of self-deception is both an involvement of the sub-system with the main system, as well as a separation of sub-system from main system. Davidson's theory is centred on just this paradox: the <u>interaction</u> between sub-system and main system is contained in the causal relation of the desire in the sub-system causing the belief in the main system, and the <u>separation</u> of the two systems is brought about by the systemic boundary which signifies a collapse in reason relation —— the sub-system fails to interact rationally with the main system. The same application of the causal link/logical breakdown is valid in addressing the fourth paradox, which is an extension of the third one.

It seems at first as if Davidson's theory constitutes a more favourable theory to adopt because it does not appeal to the problematic notion of "consciousness". Furthermore, the functional theory does not only hold for cases of "hard" self-deception but for all cases of self-deception or all cases of irrationality for that matter. Since it has a much wider scope of application than the Freudian theory it also escapes the problem of having to establish just when the irrationality is serious enough to warrant the Freudian interpretation. The functional theory need not be concerned with whether it is a case of "weak" self-deception which can be explained by one of the theories advanced in Chapter 4, or whether it is a "hard" case which needs to invoke the Freudian theory of the divided mind. The functional theory need not answer the problematic question of "Just how much counter-evidence is needed before we appeal to the unconscious?"

But, there is one severe drawback to Davidson's theory. As Pears points out:

> "However, this advantage is achieved in a way that might be found worrying. It is achieved by definition. A system's boundary is simply defined as a line across which some element in a person's psyche fails to produce its normal rational effect on the elements that control his daily life. That definition guarantees a perfect fit between the functional theory of systems and the phenomenon of irrationality that the subject is competent to avoid, but the trouble is that it seems to deprive the theory of all explanatory power." (Pears, 1984, p.84)

By appealing to a definition on which to base a theory is to rob the theory of much of its explanatory power. At least, Freud's criterion of consciousness for the division between sub-system and main system is empirically discoverable, but Davidson's criterion, it seems, leads merely

to a redescription instead of an elucidation of the original phenomenon of irrationality. In other words, the functional theory offers only the productive cause of irrationality (the wish to avoid accepting what the second-order principle points to) but not an empirically discoverable permissive cause. The explanation of irrationality is based on the failure of the sub-system to intervene rationally, but that failure is tied to the very phenomenon requiring explanation, namely the coming about of the sub-system.

However, Pears notes that, "The theory is simply not concerned with the permissive cause of irrationality in the main system. It is not intended, and it must be made clear from the start that it is not intended, as a theory about the situation in the main system that makes irrationality possible." (Pears, 1984, p.85). In other words, the functional theory does not try to oust the Freudian theory as the theory of self-deception. What it does do is to look at the other side of the coin, as a theory of self-deception, i.e. at the rational (or, rather irrational) implications rather than the psychological implications. As Davidson himself notes, "How can a person fail to put the inconsistent or incompatible beliefs together? .... It would be a mistake for me to try to answer this question in a psychologically detailed way." (Davidson, 1985, p.147). He adds that the boundaries between obviously conflicting beliefs stipulated by a theory concentrating on the rational implications are not discovered by introspection. These boundaries are rather conceptual aids to the coherent description of how inner irrationality is possible. The two emphases, of psychology on the one hand and rationality on the other, constitute one of the major differences between the two theories. However, as I have noted before, these differences are not mutually exclusive, but should be seen as complementary factors in

the theory of self-deception. According to Pears in Motivated Irrationality, the self-deceiver must "forget" or hide from himself the fact that his desire (both to avoid the painful belief and to satisfy the agent by forming the favoured belief) caused the favoured belief or caused the "rustication" of the unwelcome belief. The objection that Davidson raises to this explanation is that if the self-deceiver succeeds in hiding from himself the necessity of the contribution of the wish to the formation of the irrational belief, he is clearly self-deceived and in a state of self-deception. He is, so to speak, in a "pleasantly consistent frame of mind". (p.146). However, this pleasure of the self-deceived state is unstable and is constantly threatened by the overwhelming evidence and the agent's memory. Therefore, continued motivation is necessary and this motivation is based on the fact that the evidence points to the unwelcome belief. "If this is right, then the self-deceiver cannot afford to forget the factor that above all prompted his behaviour: the preponderance of evidence against the induced belief." (p.146). This may seem a rejection of Pears' and Bach's views, but Davidson interprets these differences as being partly due to different choices as to how to describe self-deception rather than substantive differences(4). Davidson stresses the actual process of self-deception, whereas Pears and Bach describe the state of self-deception. The differences are, therefore, complementary rather than exclusive. "To me it seems important to identify an incoherence or inconsistency in the thought of the self-deceiver; Pears and Bach are more concerned to examine the conditions of success in deceiving oneself." (p.147). Davidson, therefore, emphasizes the incoherence of self-deception which makes the irrationality clear, but then it is difficult to explain self-deception psychologically. Pears stresses and explains the actual phenomenon, but then plays down the irrationality as a result.

However, the similarities between the two theories are far greater than the alleged differences. First of all, both Davidson and Pears stress that to believe simultaneously a set of two inconsistent propositions is possible, i.e. aBp+aB-p, but it is not possible to believe the conjunction when the inconsistency is obvious, i.e. aB(p+-p), at least, impossible for a sane and rational person. Furthermore, both Davidson and Pears conceive the boundary as a dynamic aspect in self-deception. "We should not think of the boundaries as defining permanent and separate territories." (p.147). For Davidson, the sub-system falls away when the irrationality ceases — either there is no more need for self-deception or the irrationality has been pointed out to the person who, as a rational being, refrains from continuing with his irrational practice. For Pears, the boundary between sub-system and main system "shifts" when the subject is made aware of the irrationality, perhaps through psychoanalysis. What both theories, of course, rely on is the power of the sub-system over the main system. For the sub-system to effectively manipulate the main system, it needs both an internal rationality and, for the Freudian theory, an internal consciousness, which the sub-system in the functional theory already has. (The objections that can be raised to these two claims were dealt with in the last chapter, and the functional theory can be defended by the same arguments offered.)

It is now obvious how closely Davidson's theory follows the four Freudian principles mentioned at the beginning of the chapter: For Davidson certainly the mind is compartmentalized into two semi-autonomous systems, both with thoughts and beliefs. "Only by partitioning the mind does it seem possible to explain how a thought or impulse can cause another to which it bears no rational relation." (Davidson, 1982, p.303). The two systems, of course, share most of the beliefs in the person's psyche,

but certainly the two incompatible beliefs are kept apart by being relegated exclusively to different systems. The two contradictory beliefs cannot belong to the same territory —— "to erase the line between them would destroy one of the beliefs. I see no obvious reason to suppose one of the territories must be closed to consciousness, whatever exactly that means, but it is clear that the agent cannot survey the whole without erasing the boundaries." (Davidson, 1985, p.147). The "closed to consciousness" feature is one found in the theory of the early Freud, but it seems as though the view that the agent is unable to "survey the whole without erasing the boundaries" is similar to the theory which depends on the later interpretation of Freud, i.e. the postulations of "splits in the ego". As I noted in Chapter 5, the later Freud states that the mental groupings can <u>alternate</u> with their hold on consciousness (thus, not relegating one system exclusively and permanently to the unconscious) and they can remain <u>more or less</u> independent of one another. This echoes Davidson's overlapping of territories. Furthermore, Freud places the phrase "know nothing" in inverted commas when he says that these mental groupings can "know nothing" of one another. The fact that the phrase has been placed in inverted commas points to a specific employment of the meaning of "know". I am sure that it refers to the formal, logical and rational meaning of "know" and this leads us straight to the Davidsonian interpretation —— that is, there can be no <u>rational relation</u> between these two mental groupings. However, the fact that the sub-system is separated from the main system does not make it powerless. In fact, Freud has described and explained by help of experiments just how powerful the sub-system is in directing our slips of the tongue, and other forms of irrational behaviour. Therefore, even though the two mental groupings can "know nothing" of each other, there is some causal connection —— again, the Davidsonian principle in his theory of self-

deception. Similarly, for Davidson the sub-system is a directive element in our irrationality. "Being out of bounds does not make the exiled thought powerless; on the contrary, since reason has no jurisdiction across the boundary." (p.148). The two systems interact causally in that the desire in the sub-system causes the belief in the main system, but the logical relation is distorted, allowing the irrationality to exist in the main system. "What is essential is that certain thoughts and feelings of the person be conceived as interacting to produce consequences on the principles of intentional actions, these consequences then serving as causes, but not reasons, for further mental events." (Davidson, 1982, p.304). Davidson's causal relation is like Freud's "physical force" which causes the irrational belief, but reason is, of course, thwarted by the systemic boundary that has been drawn. What is needed is a certain amount of autonomy to parts of the mind.

> "The three elements of psychoanalytic theory on which I have concentrated, the partitioning of the mind, the existence of a considerable structure in each quasi-autonomous part, and non-logical causal relations between the parts; these elements combine to provide the basis for a coherent way of describing and explaining important kinds of irrationality. They also account for, and justify, Freud's mixture of standard reason explanations with causal interactions more like those of the natural sciences, interactions in which reason does not play its usual normative and rationalizing rôle." (p.304)

Davidson's theory is, therefore, not an opponent to the Freudian theory, but is rather a complementary view. However, I have not yet mentioned the last Freudian claim, i.e. the claim about unconscious mental states and events. It seems as though the notion of consciousness constitutes the damaging difference which threatens the conjunction of the two theories into <u>the</u> theory of self-deception. However, as I have discussed

earlier on, the functional theory can account for phenomena of irrationality without accepting this claim. According to the Davidsonian view, it is possible to give a plausible account of irrationality without introducing something like an unconscious piece of knowledge. Davidson does not reject the notion of the unconscious, but merely states that its introduction is unnecessary. The functional theory need not appeal to such problematic concepts like "the unconscious", but this does not mean that the functional theory denies that something like the unconscious exists. The real force of the conjunction of the two theories into one general theory is that it expands the explanatory powers of the theory.

> "If to an otherwise unobjectionable theory (the functional theory) we add the assumption of unconscious elements (the Freudian theory), the theory can only be made more acceptable, that is, capable of explaining more." (p.305)

* * * * *

In this last chapter my aim was to show how Davidson's functional theory complements rather than rivals the later Freudian theory of self-deception. First of all, the standard reason explanation for normal intentional action was given which acts as a background against which irrational behaviour can be measured. In standard reason explanations there is a desire as well as a belief about how to satisfy that desire. This belief-desire pair constitute a reason for that action. The word "reason", however, has a two-fold meaning: it refers both to the motive which causes the action as well as the logical relation of the desire-belief pair and the action they help to explain. In standard reason explanation we always have, therefore, a two fold relation: a causal and a logical or rational relation.

This being an explanation for normal intentional action, I then went on to investigate how irrational action deviates from this, where the breakdown occurs. As Davidson noted, in order to explain irrational behaviour it is necessary to see it against a background of rationality and explanations of irrational behaviour must, therefore, have a rational core — the Plato Principle ——if we are to make any sense of irrational behaviour at all. For Davidson then, explanations of irrational behaviour entail a desire which acts as a cause for a belief, but which is not a justified cause, i.e. the logical relation is distorted. Davidson distinguishes between two forms of irrational behaviour. Firstly, there is behaviour that seems irrational to us — external irrationality — but need not necessarily involve the agent's going against his own second-order principle. This is especially the case in akrasia and in "weak" self-deception in which the agent has reasons both for and against a certain belief or course of action, but then, basing his judgment on only part of the evidence, he chooses that belief or course of action which he desires, but which goes against the better reasons he has for the other option. This may seem irrational to us, but unless the agent accepts a second-order principle, the belief or action need not be internally irrational. And it is especially this second case of irrationality, internal irrationality, that interests us. This is, of course, the area of "hard" self-deception in which the evidence is so overwhelmingly in favour of the unwelcome belief that the agent, as a rational person, cannot help but form that belief.

It was necessary at this stage to distinguish between the various implications of the use of "reason". Pears notes that a desire can act as a reason for an action or a belief but then it acts as an irrational reason. A desire can be a reason for having a belief but not a reason

for the belief itself. The self-deceiver brings about his own "reason" for the belief itself and this is the point in the sequence at which irrationality enters. Davidson employs "reason" as being synonomous with "rational", but this led to criticism when he states that in irrational behaviour a mental state can be a mental cause of another mental state but not a reason for it. Pears rephrases this dictum to read that in irrationality there is a mental cause that is not a rational reason for what it causes.

In order to explain how "hard" self-deception is possible, Davidson bases his theory on the four Freudian principles. He states that in an interpersonal situation there can be a mental state which causes another mental state but is not a rational reason for it. The same is possible within an intrapersonal situation, i.e. within one mind. But it is only by partioning the mind that it seems possible to explain how a thought can cause another to which it bears no rational relation. The mental parts are conceived, like in Freud's theory, as semi-autonomous agents, but what is essential is that there is interaction between certain thoughts and feelings of the person, which produces consequences on the principles of intentional actions. These consequences serve as causes, but not rational reasons, for further mental events. And if some mental events act as causes for some other mental event in the same mind, a degree of autonomy to parts of the mind is necessary.

The autonomy is brought about by the drawing of the boundary which keeps the inconsistent beliefs apart and which makes it possible for the subsystem to fail to interact rationally in the main system. The drawing of the boundary is in itself an irrational step for it is caused by the desire to avoid following the rational path in which the second-order

principle directs the agent, but the drawing of the boundary is not a rational reason for allowing the agent to practise his self-deception unhindered. As Davidson notes, "Nothing can be viewed as a good reason for failing to reason according to one's best standards of rationality." (p.148). There is, of course, the distinction then of a reason for ignoring the second-order principle and a reason for thinking that the principle is no longer good or that he need not abide by it. There can be a reason for the former, but not for the latter. The agent has a motive or desire that is the cause of his behaviour and it may be a "good" reason under certain circumstances — Emma may reason that if she does not ignore the second-order principle, she will have a nervous breakdown. This case is describable in the language of rationality in the broad sense, but there cannot be a reason for going against (as opposed to ignoring) the over-arching principle of rationality. Irrationality is, therefore, explained in terms of sub-systems perfectly coherent in themselves, but disjoined from the main system so as to permit the irrationality to flourish there.

This notion ties up with Freud's description of irrationality in terms of some self-division. It is especially the later description of these divisions as "splits in the ego" and which employ the notion of the conscious in, what Pears terms, a "functional" sense that link up very closely with that of Davidson. The two theories are, therefore, complementary to each other. Davidson's theory concentrates on the rational aspect of a theory of self-deception and Freud's on the psychological aspects. The balance between the two is delicate because, as Davidson notes, to stress the one is to underplay the other. How we interpret self-deception will depend on our choice of description.

Notes

1. The other extremist interpretation is founded on the Medea Principle in which an action, although intentional, is not in itself an action for which the agent can be held responsible. If the agent is to blame, it is not for what he did, but because he did not resist with greater resolve. What the agent did had a reason ——an overwhelming passion ——but the reason was not his and, therefore, not truly intentional. However, this case of irrationality brought about by "outside" stronger forces is not the one that creates conceptual difficulty and is, therefore, not discussed in any detail.

2. Tim can, of course, retrieve the letter and still communicate his disappointment in other ways ——e.g. by acting coolly towards Eric, not returning his telephone calls, hinting at his disappointment to a third person, etc. But I have deliberately simplified the example by excluding these options and so tailored the example to illustrate the principles of Davidon's theory.

3. Baier (1985) accuses Davidson of freeing the akrates not only from the charge of inconsistency, but also from the charge of irrationality, if he disavows, or simply doesn't avow, the second-order principle. Although Pears (1985) agrees with Baier in stressing the importance of the appropriate second-order principle, he feels that she is harsh in her cirticism, for Davidson does not deny its importance. Davidson's point is only that the difficult thing is to explain internal irrationality, but that, if it becomes clear that an agent is not being internally irrational because he does not accept the appropriate second-order principle and is, nevertheless, externally irrational, that difficulty vanishes. There is no suggestion by Davidson that there is nothing wrong with external irrationality, or even that it does not need to be explained. The point is simply that the special difficulty of internal irrationality does not arise in this kind of case, because there is no internal irrationality.

4. In Chapter 4 I discussed Bach's view. This view advocates that the self-deceiver cannot actually believe in the weight of the contrary evidence. Bach shares Pears' view that the actual attainment of the state of self-deception cannot co-exist with the original motivation of that state ——the conflict is just too great.

Bibliography

Abelson, R. (1977). Persons - A Study in Philosophical Psychology. London: MacMillan Press, 1977.

Audi, R. (1976). "Epistemic Disavowals and Self-Deception" in Personalist, vol.57, 1976.

Audi, R. (1982). "Self-Deception, Action, and Will" in Erkenntnis vol.18, 1982.

Audi, R. (1984). Critical Review - M.R. Haight, "A Study of Self-Deception" in NoÛs, vol.18, 1984.

Bach, K. (1980). "An Analysis of Self-Deception" in Philosophy and Phenomenological Research, vol.41, 1980-81.

Bach, K. (1984). "More on Self-Deception: Reply to Hellman" in Philosophy and Phenomenological Research, vol.45, 1984-85.

Baier, A.C. (1985). "Rhyme and Reason: Reflections on Davidson's Version of Having Reasons" in Actions and Events: Perspectives on the Philosophy of Donald Davidson - E. Le Pore and B.P. McLaughlin (ed.) Oxford; New York: Basil Blackwell, 1985.

Benn, S.I. and G.W. Mortimore (ed.) (1976). Rationality and the Social Sciences. London: Routledge and Kegan Paul, 1976.

Bennett, J. (1967). Rationality - An Essay Towards an Analysis. London: Routledge and Kegan Paul, 1967.

Brooks, D.H.M. (1986). The Unity of the Mind. Unpublished doctoral thesis, University of the Witwatersrand, 1986.

Canfield, J.V. and D.F. Gustavson (1962). "Self-Deception" in Analysis, vol.23, 1962.

Cavell, M. (1986). "Metaphor, Dreamwork and Irrationality" in Truth and Interpretation - Perspectives on the Philosophy of Donald Davidson. E. Le Pore (ed.). Oxford; New York: Basil Blackwell, 1986.

Champlin, T.S. (1976). "Double Deception" in Mind, vol.85, 1976.

Champlin, T.S. (1977). "Self-Deception: A Reflexive Dilemma" in Philosophy, vol.52, 1977.

Champlin, T.S. (1979). "Self-Deception: A Problem about Autobiography" in The Aristotelian Society, suppl. vol.53, 1979.

Cohen, L.J. (1986). The Dialogue of Reason. Oxford: Clarendon Press, 1986.

Davidson, D. (1963). "Actions, Reasons and Causes" repr. in his Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, D. (1970a). "How is Weakness of the Will Possible? repr. in his Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, D. (1970b). "Mental Events", repr. in his Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, D. (1971). "Agency", repr. in his Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, D. (1978). "Intending" repr. in Essays on Actions and Events. Oxford: Clarendon Press, 1980.

Davidson, D. (1982). "Paradoxes of Irrationality" in Philosophical Essays on Freud, R. Wollheim and J. Hopkins (ed.). Cambridge: Cambridge University Press, 1982.

Davidson, D. (1985). "Deception and Division" in Actions and Events: Perspectives on the Philosophy of Donald Davidson. E. Le Pore and B.P. McLaughlin (ed.). Oxford; New York: Basil Blackwell, 1985.

Davidson, D. (1985b). "Reply to Grice and Baker" in Essays on Davidson - Actions and Events. B. Vermazen and M.B. Hintikka (ed.). Oxford: Clarendon Press, 1985.

Demos, R. (1960). "Lying to Oneself" in The Journal of Philosophy, vol.57, 1960.

De Sousa, R.B. (1970). "Review of Self-Deception" in Inquiry, vol.13, 1970.

Elster, J. (1979). Ulysses and the Sirens. Cambridge: Maison des Sciences de l'Homme and Cambridge University Press, 1979.
Elster, J. (1985). Sour Grapes - Studies in the Subversion of Rationality. Cambridge: Maison des Sciences de l'Homme and Cambridge University Press, 1985.
Elster, J. (1985b). "The Nature and Scope of Rational-Choice Explanation" in Actions and Events: Perspectives on the Philosophy of Donald Davidson, E. le Pore and B.P. McLaughlin (ed.). Oxford; New York: Basil Blackwell, 1985.
Factor, R.L. (1977). "Self-Deception and the Functionalist Theory of Mental Processes" in Personalist, vol.58, April 1977.
Fingarette, H. (1969). Self-Deception. London: Routledge and Kegan Paul, 1969.
Fingarette, H. (1982). "Self-Deception and the 'Splitting of the Ego'" in Philosophical Essays on Freud, R. Wollheim and J. Hopkins (ed.) Cambridge: Cambridge University Press, 1982.
Foss, J. (1980). "Rethinking Self-Deception" in American Philosophical Quarterly, vol.17, no.3, July 1980.
Freud, S. (1949). Complete Introductory Lectures on Psychoanalysis, trans. J. Strachey. London: George Allen and Unwin, 1971.
Gardiner, P. (1970). "Error, Faith and Self-Deception" in Proceedings of the Aristotelian Society, vol.70, 1969/70.
Gide, A. (1919). The Pastoral Symphony in "Two Symphonies", trans. D.Bussy. New York: Vintage Books, 1977.
Grice, P. and J. Baker. (1985). "Davidson on 'Weakness of the Will'" in Essays on Davidson - Actions and Events, B. Vermazen and M.B. Hintikka (ed.). Oxford: Clarendon Press, 1985.
Goleman, D. (1987). "Who Are You Kidding?" in Psychology Today, March 1987.
Haight, M.R. (1980). A Study of Self-Deception. Sussex: The Harvester Press; N.J.: Humanities Press, 1980.
Hamlyn, D.W. (1971). "Unconscious Intentions" in Philosophy, vol.46, 1971.
Hamlyn, D.W. (1971a). "Self-Deception" in Proceedings of the Aristotelian Society, vol.45, 1971.
Hanson, K. (1986). The Self Imagined. New York: Routledge and Kegan Paul, 1986.
Hellman, N. (1983). "Bach on Self Deception" in Philosophy and Phenomenological Research, vol.44, 1983-84.
Hollis, M. and S. Lukes (ed.)(1982). Rationality and Relativism. Oxford: Basil Blackwell, 1982.
Hurley, S.L. (1985). "Conflict, Akrasia and Cognitivism" in Proceedings of the Aristotelian Society, vol.86, 1985-86.
Kekes, J. (1976). A Justification of Rationality. New York: State University of New York Press, 1976.
Kipp, D. (1980). "On Self-Deception" in The Philosophical Quarterly, vol.30, 1980.
Kittay, E.F. (1982). "On Hypocrisy" in Metaphilosophy, vol.13, 1982.
Martin, M.W. (1979). "Self-Deception, Self-Pretence, and Emotional Detachment" in Mind, vol.88, 1979.
Martin, M.W. (1986). Self-Deception and Morality. Lawrence: University Press of Kansas, 1986.
Mele, A.R. (1982). "'Self-Deception, Action and Will': Comments" in Erkenntnis, vol.18, 1982.
Mele, A.R. (1983). "Self-Deception" in The Philosophical Quarterly, vol.33, no.133, 1983.
Mele, A.R. (1984). "Pears on Akrasia, and Defeated Intentions" in Philosophia, vol.14, August 1984.
Miri, M. (1973). "Self-Deception" in Philosophy and Phenomenological Research, vol.34, 1973-74.

Nye, R.D. (1975). Three Views of Man. Monterey, Calif.: Wadsworth Publishing Co., 1975.
Palmer, A. (1979a). "Characterising Self-Deception" in Mind, vol.88, 1979.
Palmer, A. (1979b). "Self-Deception: A Problem About Autobiography" in The Aristotelian Society, suppl. vol.53, 1979.
Paluch, S. (1967). "Self-Deception" in Inquiry, vol.10, 1967.
Parfit, D. (1984). Reasons and Persons New York: Oxford University Press, 1984.
Peacocke, C. (1985). "Intention and Akrasia" in Essays on Davidson - Actions and Events, B. Vermazen and M.B. Hintikka (ed.). Oxford: Clarendon Press, 1985.
Pears, D.F. (1974). "Freud, Sartre and Self-Deception" in Freud, R. Wollheim (ed.) N.Y.: Doubleday, 1974.
Pears, D.F. (1982a). "Motivated Irrationality, Freudian Theory and Cognitive Dissonance" in Philosophical Essays on Freud, R. Wollheim and J. Hopkins (ed.). Cambridge: Cambridge University Press, 1982.
Pears, D.F. (1982b). "Motivated Irrationality" in The Aristotelian Society, suppl. vol.56, 1982.
Pears, D.F. (1984). Motivated Irrationality. New York: Oxford University Press, 1984.
Pears, D.F. (1985). "Reply to Annette Baier: Rhyme and Reasons" in Actions and Events - Perspectives on the Philosophy of Donald Davidson, E. Le Pore and B.P. McLaughlin (ed.). Oxford; New York: Basil Blackwell, 1985.
Penelhum, T. (1964). "Symposium: Pleasure and Falsity" in American Philosophical Quarterly, vol.I, no.2, April 1964.
Phillips, D.Z. (1981). "Bad Faith and Sartre's Waiter" in Philosophy, vol.56, 1981.
Priest, G. (1986). "Contradiction, Belief and Rationality" in Proceedings of the Aristotelian Society, vol.86, 1985-86.
Pugmire, D. (1969). "'Strong' Self-Deception" in Inquiry, vol.12, 1969.
Pugmire, D. (1982). "Motivated Irrationality" in The Aristotelian Society, suppl. vol.56, 1982.
Radden, J. (1984). "Defining Self-Deception" in Dialogue, vol.23, 1984.
Russel, J.M. (1978). "Saying, Feeling and Self-Deception" in Behaviourism, vol.6, 1978.
Saarinen, E. (1985). "Davidson and Sartre" in Actions and Events - Perspectives on the Philosophy of Donald Davidson, E. Le Pore and B.P. McLaughlin (ed.). Oxford; New York: Basil Blackwell, 1985.
Sartre, J.P. (1958). Being and Nothingness, trans. H.E. Barnes. London: Methuen & Co.Ltd., reprint 1984.
Sartre, J.P. (1982). "Mauvaise foi and the Unconscious" in Philosophical Essays on Freud, R. Wollheim and J. Hopkins (ed.). Cambridge: Cambridge University Press, 1982.
Saunders, J.T. (1975). "The Paradox of Self-Deception" in Philosophy and Phenomenological Research, vol.35, 1974-75.
Solomon, R.C. (1978). "Self-Deceptive Emotions" in The Journal of Philosophy, vol.75, no.7, July 1978.
Siegler, F.A. (1962). "Demos on Lying to Oneself" in The Journal of Philosophy, vol.59, 1962.
Szabados, B. (1973). "Wishful Thinking and Self-Deception" in Analysis, vol.33, 1972-73.
Szabados, B. (1974). "Self-Deception" in Canadian Journal of Philosophy, vol.4, 1974.
Thalberg, I. (1985). "Questions About Motivational Strength" in Actions and Events - Perspectives on the Philosophy of Donald Davidson, E. Le Pore and B.P. McLaughlin (ed.). Oxford; New York: Basil Blackwell, 1985.

Watson, G. (1977). "Skepticism about Weakness of Will" in The Philsophical Review, vol.86, 1977.