S E L F - D E C E P T I O N

Bernard Clarke

Philosophy Dept. PhD.

February 1992

UNIVERSITY WARWICK

Contents

Introduction

> in the midst of a difficult problem, the Danish physicist Niels Bohr was overheard to say, "how wonderful that we've met with a paradox. Now we have hope of making some progress." (Von Oech [1983], p118)

There is a reflexive paradox (or set of paradoxes) associated with self-deception, and a variety of theories have been proposed in response, to explain self-deception.

The study of reflexive paradoxes has been fruitful in the history of philosophy. Such a paradox may appear to be no more than a minor puzzle, which we will easily be able to mop up after having formulated solutions to more major problems. Sometimes the minor puzzle turns out to be surprisingly resistant to our "mopping up" operations; it may force us to re-think our major theories. For example the "truth-teller" paradox and other paradoxes of self-reference have been viewed initially as minor puzzles, while later on they have provoked major theories, e.g. theories of truth; in mathematics, Godel's theorem.

In subsequent chapters I will discuss the paradox of self-deception and the theories which have been developed in response to the paradox. I shall emphasise what I take to be the good points and the bad points of each theory. These theories are not all in conflict; some of them are complementary, one supplying what is lacking in another. But there is no one theory with which I wholeheartedly agree; some of them I find undeveloped (and so I am not sure what there is with which to agree or disagree). Where I feel that someone's discussion of self-deception has alluded to a theory rather than actually presented that theory, I have tried to develop the allusion to a stage at which we can find out to what sort of theory they are alluding. This means that (borrowing an expression from the plant breeders) the theories which I discuss are "pure strains", which I have bred from the "hybrids" presented by other writers. This is not a criticism of the other writers: I think that "hybrids" is the right way to go if we wish to explain self-deception. The point of the "pure strains" is to sharpen up our understanding of what is at issue in the discussions about self-deception.

Dissatisfaction with the available theories provoked me to try to develop my own theory about self-deception. Several chapters are devoted to challenging what I take to be false assumptions which lead to the paradox. I then present my own account of self-deception and in the final chapter discuss the

consequences of that account. In particular, I argue that often discussions in epistemology are guided (or rather misguided) by a collection of metaphors which have inhabited philosophy from ancient times. These "dead" metaphors are still very active and effectively prevent us from developing some new theories which we need.

Before entering into the detailed discussions contained in the following chapters, I want to briefly summarise their content. The aim is to map out my line of argument so that the reader knows where the discussion is heading and "what I am driving at".

Firstly, lets present the paradox (or paradoxes) of self-deception. There are many formulations of the paradox, but a typical version goes like this: the self-deceiver must know what he (or she) is up to, otherwise he is not really self-deceived but merely mistaken. Self-deception is different from mistake. If he knows what he is up to then he is not really deceived. For someone who knows he is self-deceived also knows that the belief he has as a result of the deception is false. But if someone knows that what he "believes" is false, then he does not really believe it. At most he may pretend to believe it. Therefore 'self-deception' is always a misnomer: either the alleged "self-deception" turns out to be merely a mistake, or else it turns out not to be genuine deception. If someone

is genuinely self-deceived then he must believe something which he knows to be false, but this, we have shown, is impossible.

'Self-deceived' and other expressions which mean the same thing (such as 'kidding himself', 'fooling herself' and 'being dishonest with oneself') can be found in a vast number of works about people, including both fiction (novels, plays) and non-fiction (history, sociology, psychology). Reading some of these works makes it apparent that when people refer to self-deception, they intend to distinguish it from "honest mistakes", and from "honest pretences" (e.g. joking, playacting); they also do not regard self-deception as equivalent to any of the varieties of dishonesty and insincerity which have other people as their intended audience. Self-deception is not a lie to other people: it is sometimes described as lying to oneself.

One response to the paradox is to agree that it shows self-deception to be impossible, therefore there is no such thing. I call this the "No Such Thing" theory of self-deception.

Another response is the Schism theory. Schism theories point out that the paradox arises because it is reflexive: the deceiver deceives himself. Schism theories propose that the self-deceiver is divided, with one part deceiving the other part. This situation is supposed to explain without paradox the

behaviour which causes other people to describe someone as self-deceived.

I reject Schism theories; where there is independent evidence of a schism we do not find self-deception, and where we find alleged self-deception there is no independent evidence of a schism. Therefore the main strength of Schism theories is the claim that there is no other way to explain self-deception, and so we have to postulate that a person is divided, when we call them self-deceived. The Schism theory is vulnerable to competition from other theories of self-deception. And there are a few other theories.

Dissociation theories (or "Disconnection theories" - I use the two descriptions interchangeably) argue that the self-deceiver somehow disconnects the knowledge he has from the false beliefs he has, or manages to "think" something which he "believes" to be false, or disconnects true beliefs from actions or emotions - there is disagreement among the different theories about what is disconnected from what, and where the disconnection happens. How it happens also goes unexplained, and some of the terminology used strikes me as in need of more explanation.

I also have a suspicion that perhaps Disconnection theories are saying, in a roundabout and imprecise way, something which is

just as paradoxical as the original "the self-deceiver must know what he is up to": "thinking something you do not believe" seems to me rather like "believing something you know to be false", when it is used in the way Disconnection theories do, i.e. as something which is going be substituted for "believing something you know to be false". For the "thinking" is going to have to do all the work of what was called "believing" in the original formulation of the paradox.

Role theories of self-deception offer an explanation of the disconnection: the self-deceiver does not believe that p (where 'p' is some expression which we use to identify a belief), but he adopts the role of someone who does believe it; and to do so, he uses p as he would use a belief: p is given the role of a belief.

Role theories are open to the objection that what they describe is not self-deception since they do not describe genuine belief, only someone pretending to believe. The Role theory is an unparadoxical account, but only because it abandons the problem: the "self-deceiver" is not deceived.

I think that Role theories make some very good points, and the objection can be answered in a satisfactory way, when the good points of Role theories are put together with the good points of the Negligence theory of self-deception. The Negligence

theory claims that self-deception is epistemic negligence: the self-deceiver does not do the things which are needed to gain truth: self-deception is not a positive act of seeking falsehood but the negative act of not seeking truth (or of not trying hard enough when seeking truth).

Negligence theories are vulnerable to the claim which gives rise to the paradox: "the self-deceiver must know what he is up to". He needs to know in order to guide the strategy of being self-deceived: in order to steer away from evidence which would threaten to destroy his preferred beliefs. Someone who is merely negligent has no control over his beliefs: but a self-deceiver, typically, is not content to have any beliefs which happen to occur to him: he has preferences as to which beliefs are acceptable to him.

I claim that the paradox-generating claim can be refuted. The self deceiver does not need to "know what he is up to", indeed he must not know what he is up to (for that way lies paradox). The self-deceiver needs to use a theory in order to guide the strategy of self-deception, but it does not need to be a true theory: for some purposes a false theory can be just as effective as a true theory, or more effective. So the self-deceiver need not "know what he is up to", but he may need an effective (though false) theory. No theory at all will be

needed if the "strategy" of self-deception is guided by something else, something which is not a theory at all.

Self-deception is a way of managing self-misunderstanding. Lack of knowledge does not mean that the self-deceiver is merely mistaken. Negligence is not a mere mistake: someone who is negligent can be held responsible for being negligent.

Having rejected the claim which generates the paradox, we are in a position to construct an effective explanation of self-deception. The self-deceiver is able to "know what he is up to", but unwilling to exercise this ability. He is unwilling to do so, not merely too lazy to do so: for if he were merely too lazy then any beliefs would do, true or false. Whereas the deceiver displays a preference for some beliefs over others.

Instead of using knowledge, the self-deceiver uses a false theory, or some other means which is not a theory, to "guide the strategy". But, it can be objected, if the theory is false then it will not enable the self-deceiver to predict and so avoid evidence which can refute, and so destroy, his favoured belief. My reply is two-fold. Firstly, evidence is not a natural product which we might just accidentally bump into: evidence is made. Normally to say that evidence is made (or, even more derogatory, "fabricated") means that it is counterfeit: it is not genuine evidence. This is not what I

mean. My claim is that all evidence (including "genuine" evidence) is made. There are processes which create evidence. They are related to the processes which create beliefs. The self-deceiver's belief is not vulnerable to destruction by evidence because although he is able to construct that evidence, he does not do so.

Secondly, evidence is not coercive: it does not compel us to believe the conclusions it makes evident. There is a set of metaphors which encourage us to think that evidence is coercive (it is "forceful", "compelling" etc). But these metaphors should be resisted. Beliefs are not "based" on evidence, for they are not "based" at all. They are invented, and tested (if at all) in practice.

Everyone agrees that our falsehoods are invented. Not everyone agrees that our truths are also invented. But they are. We get both truths and falsehoods through the same type of process. The self-deceiver is not doing something special and different from the truth-seeker when he invents beliefs. But he invents his beliefs not to gain truths but for some other purpose. The self-deceiver might even invent a true belief and still be self-deceived, for self-deception is characterised by the purpose with which it is done: someone could deceive himself into believing something which is true. He would nonetheless be a self-deceiver because truth was not his goal.

To sum up: I claim that the available theories do not provide a satisfactory explanation of self-deception. I offer a theory which I take to provide a satisfactory explanation, but there is a cost attached to it: we have to give up some assumptions which, it seems to me, are deep-rooted in the terminology of epistemology.

We must give up the idea that evidence is a generator of belief; the justification of belief comes after we adopt the belief, not before: "the proof of the pudding is in the eating". Evidence, and truths, are not found but made. We "find" that our truths are true by trying them. If they "work", then we cannot distinguish them from truths. The only sort of thing which in all circumstances is indistinguishable from a truth, is a truth.

Plato, in the Republic (382d), writes:

> we don't know the truth about the past but we can invent a fiction as like it as may be. (Plato [1974] p138)

Plato uses the word 'pseudos', translated as 'fiction' in the quotation above: 'pseudos' can also mean 'falsehood'. My footnote to Plato is that if the fiction is exactly like the truth, then it is true. We say what things are by saying what they are like.

> How could something originate in its antithesis? Truth in error, for example? Or will to truth in will to deception? ... Such origination is impossible; he who dreams of it is a fool, indeed worse than a fool: the things of highest value must have another origin of their own - they cannot be derivable from this transitory, seductive, deceptive, mean little world, from this confusion of desire and illusion. (Nietzsche [1973] p15)

Nietzsche writes this in irony . But he also writes:

> a philosopher: alas, a creature which often runs away from itself, is often afraid of itself - but which is too inquisitive not to keep 'coming to itself' again. (Nietzsche [1973] p198).

The claim that "truth is what works" has been derided by people who did not bother to distinguish the work for which we want truth from the work performed by other sorts of interpretive instrument - such as the self-deceiver's belief. A study of the "work" performed would enable us to define truth by the way our truths are grounded in the way we live, in what Wittgenstein called our "forms of life". It would provide more than a dry definition: it would give us insights into ourselves, our enquiries, and our need for truth.

This is not the sort of insight which the discipline of psychology seeks: my thesis is not about scientific laws which explain how one sort of thing - people - work, or what they are: it is more like the "user guide" which explains, in layman's language, how to use something. Even though, from a technical point of view, it may misrepresent the inner workings of the thing, the thing behaves "just as though it were true".

This thesis is not intended to be an essay in psychology - not even in "folk psychology", if that is meant to be some kind of competitor to the academic discipline of psychology. The thesis has more to do with making our theories of truth and evidence workable: an essay in epistemology, not psychology. Some epistemic theories help us to explain self-deception while others hinder. We should drop the ones which hinder us.

The self-deceiver is not engaged in a project very different from that of someone seeking truth. Both seek an instrument that "works", but their aims in seeking that instrument differ, and so the work to be done also differs. Epistemology has a long history of attempting to provide a methodology for enquiry: for seeking truth or, as I would say, for making truths. The need for such a methodology indicates that there are other ways of proceeding, which we can adopt at will. Enquiry is only one among many procedures for making the instruments which we use for processing information. The

assumption that it is the only one contributes to making the "paradox" of self-deception insoluble. The instruments are interpretations - theories, for example. The processes by which we construct these instruments are processes of interpretation. Later on I shall have a good deal to say about interpretations (see my chapter on "Radical Interpretation theories" about self-deception) and about processes (see my discussion of the "Process theory" about self-deception, in the chapter titled, "The Production Process").

In the next chapter I offer a classification of the theories about self-deception.

A Note On Terminology

I have generally used the words 'he', 'him' and 'his' as generic expressions rather than as gender-specific designations. The generic expressions can be translated into the gender-neutral but cumbersome expressions - 'he or she', 'him or her', 'his or hers' - without contradicting the sense of the text, though the result is rather difficult to read. No sexism is intended and I certainly do not wish to imply that women are any less likely or any more likely to be self-deceived than are men.

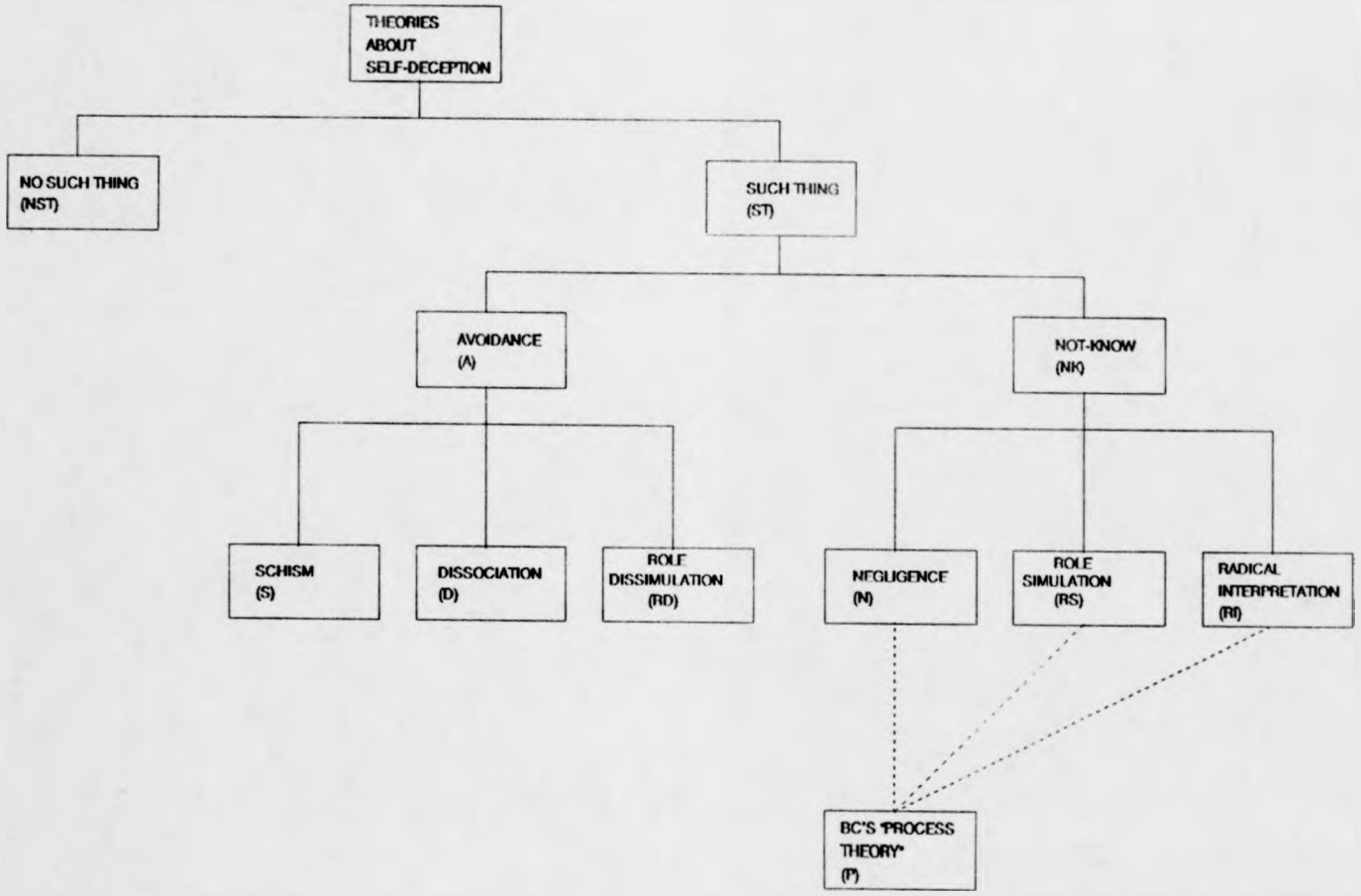
A Classification Of Theories

I suggested that the theories of self-deception are usually "hybrids" of "pure strains". In this chapter I classify these pure strains. I also outline the basis for classifying theories in this way. This is only one way of classifying them; but it is one which is well adapted to my aim in this thesis.

Diagram 1, on the following page, summarises the classification. It shows a hierarchy of types of theory. The first branching in the hierarchy is the division between those theories which claim that there is such a thing as self-deception and "No-such-thing" (NST) theories, which argue that 'self-deception', taken literally, is a misnomer. They claim that while there is something which we call "self-deception", the expression 'self-deception' misdescribes it.

NST theories use the paradoxes of self-deception to corroborate their claim. They therefore fail if any other theory provides a solution to the paradoxes.

CLASSIFICATION OF THEORIES



Other theories - lets call them Such-Thing (ST) theories by contrast with No-Such-Thing theories - are divided up according to the way they aim to solve the paradoxes.

To begin with, allegedly "the self-deceiver must know what he is up to". Some theories deny this. I call these "Not-Know" (NK) theories. Other theories are designed to show how self-deception is possible even when the self-deceiver knows what he is up to. They do so by suggesting that the self-deceiver's knowledge is disabled or segregated in some way so that it cannot inhibit the self-deception. These I call "Avoidance theories".

Avoidance theories come in three varieties:

- Schism theories
- Dissociation (or Disconnection) theories
- Role theories.

Schism theories argue that the self-deceiver is divided, one part being the deceiver and another part being the deceived.

Dissociation theories do not go so far as proposing a schism in the self-deceiver; they argue that:

- the effects of the self-deceiver's knowledge are altered by disconnecting the knowledge from other things - things such as action, emotion, "thinking" etc

- or that the self-deceiver "knows in one sense and not in another".

Role theorists argue that, by playing a role, the self-deceiver disables the knowledge of what he is up to.

Role theories also appear under the heading of "Not-Know theories" - see below for an outline of how Not-Know Role theories differ from Avoidance Role theories.

Not-know theories also come in three varieties:

- Negligence theories

- Role theories

- Radical Interpretation theories.

Negligence theories argue that self-deception can be explained as a form of epistemic negligence. They claim that self-deception is achieved by not doing the things which would lead to true beliefs and prevent false beliefs. Since someone's

negligence does not exculpate them, we can explain how the self-deceiver is responsible for being deceived even though he does not "know what he is up to".

Role theories describe self-deception as the adoption of a role: the role can guide the strategy of self-deception even though the self-deceiver does not "know what he is up to".

Role theories come in two flavours: some of them are Avoidance theories, according to which one may be self-deceived by adopting a role in order to mask or draw attention away from what one really believes or feels. The role may be used to defer making use of a belief (or to defer the loss of a belief), for an indefinitely long time and perhaps forever. I call theories of this sort "Role Dissimulation" theories, and distinguish them from "Role Simulation" theories. Role Simulation theories are Not-Know theories. They do not make the claim that adopting a role masks or defers beliefs. There need be no "genuine" belief lurking behind or beneath the role: the role need not be a mask. Role simulation theories claim that adopting a role in a thorough and unrestricted way results in genuine beliefs, not pretences.

Radical Interpretation theories argue that self-deception is achieved by the self-deceiver creating (rather than "adopting") false beliefs, and that knowledge is not needed to guide self-

deception - a false theory can be as effective an instrument of control as a true one.

I describe my own account of self-deception as a "Process Theory". The process theory is intended to explain self-deception by giving an account of the processes by which beliefs are produced. I claim that the process theory can be used to spell out what makes the other theories appealing.

That is not to say that the other theories are correct. The Process theory is not neutral with regard to the other theories. It is a Not-know theory: part of the claim is that Avoidance theories are false, and in any case the Process theory makes them redundant - because it is a better theory.

Other Not-know theories are limited and insufficiently spelled-out rather than incorrect.

There is a further distinction which cuts across the classification which I have proposed. This is the distinction between cognitive and non-cognitive theories of self-deception. Supporters of non-cognitive theories argue that many accounts of self-deception put too much emphasis on the cognitive issues - e.g. knowledge and belief - and that more discussion of the non-cognitive issues - such as volition, action, and emotion - is needed.

I agree with this up to a point. The point at which I cease to agree with it is when the cognitive issues disappear from the discussion altogether. The paradoxes of self-deception are cognitive paradoxes. If we cease to discuss the cognitive issues then the paradoxes will seem to be "dissolved". But that is because we have simply ceased to talk about them. That does not solve the paradoxes, it ignores them. So an exclusively non-cognitive approach would not help unless we could lead exclusively non-cognitive lives. This seems unlikely and so we need to discuss both: the cognitive issues and the non-cognitive.

In this chapter I have outlined the basis of my classification of the theories and named the major "pure strains". I will discuss each of the pure strains in more detail. Before doing so I have some remarks to make about the use of examples in discussions of self-deception. That is the subject of the next chapter.

A Problem About Examples

Many discussions of self-deception start by giving an example of self-deception and then go on to discuss it. I have not taken that approach because I think it is very difficult and perhaps impossible to describe an example in a way which is neutral with regard to the various theories about self-deception.

Any example we select will encounter problems such as the following.

1. Anyone who agrees with the alleged "self-deceiver's" belief is likely to argue that the example is not an instance of self-deception - in self-defence. For otherwise they are conceding that in all probability they are self-deceived too. Perhaps this is why the examples given are almost invariably examples of isolated individuals deceiving themselves, never of large groups of people. An exception to this rule may be Marx's discussion of "false consciousness", which I take to imply that large groups of people are self-deceived in the same way and about the same thing (and other large groups are deceived though not self-deceived by false consciousness). Another exception may be Nietzsche's claim that "the will to truth" is

a "will to deception". I discuss both Marx and Nietzsche later on.

This problem is ameliorated if we suggest (as I do) that someone may acquire true beliefs in a self-deceptive way. They will then have deceived themselves into believing something true. This is still not going to allow us to create an example that we can all agree on if the "self-deceptive way" is characterised in such a manner that it fits the behaviour of the people whose agreement we are seeking.

2. No example of self-deception can gain agreement anyway, since if we describe it as self-deception then No-Such-Thing theorists will argue that it is better described as something other than self-deception. So any example we choose will fail to be theory-neutral.

3. What we can do is to set up an example and invite the No-Such-Thing theorist to knock it down. But this is problematic too. If we invent an example, then the NST theorist can argue that it is not realistic or life-like. If we use an example taken from life then the NST theorist can argue (probably correctly) that we do not know enough about the alleged self-deceiver's motives, or that we have not spelled out enough of the circumstantial details which, if they were added in to the example, might alter our claim that it is an example of self-

deception. If we borrow an example from literature, e.g. Karenin from Tolstoy's Anna Karenina, then we borrow the theory which the novelist has built into the example, and we cannot add in all the circumstantial detail which we may think relevant, because the novel does not supply it.

4. Suppose that we ignore all these problems and plough on with the attempt to give an example of self-deception. The example we give will be constructed to rule out explanations other than the claim that this is self-deception. This sort of construction is provided all the time in discussions of self-deception.

For example, the example may be that of a doctor dying of a terminal illness. The illness is her speciality so she, of all people, should know that she is dying. She demonstrably knows too much to be mistaken: and yet she behaves as if unaware that she is dying, and asserts that she is as fit as a fiddle. She might be denying the facts in order to deceive others, knowing that their attitude towards her will change, perhaps. But we can rule this out by constructing the example so that she is surrounded by sensible people whom she trusts will not alter their attitude towards her. We can rule out the motive by characterising her as never having cared what other people thought. Ruling out the motive to deceive others restricts the example to something which seems more like self-deception. For

although she has no motive to deceive others, she may have plenty of reasons for deceiving herself. She wishes to avoid the painful emotions and moods of black depression which she would suffer if she believed that she was dying. She wants to go on working and the moods and emotions would prevent her doing so ... and so on.

Once again the theory we use is going to impinge upon our example. For a supporter of Dissociation theories can argue that she does not need to "avoid the truth" in order to achieve her goals: she only needs to avoid "the glaring truth". She can simply pretend that she is not dying, and that will be sufficient to evoke more comfortable emotions and moods. At this point the NST theorist may well point out that if she is merely pretending, then she is not self-deceived.

Someone who supports a Schism theory of self-deception may argue that mere pretence may not be sufficient to achieve her aims. For how can the pretence dispel the unwelcome moods and emotions if her belief is still present and evoking them? Even if the pretence works in this case, there may be other examples where a pretence will not be sufficient to achieve the person's goals, and will not be sufficient to explain what is going on. At this point the Schism theorist is likely to start constructing other examples, and challenging the Dissociation theorist to explain them.

5. The construction of examples is inextricably bound up with the theory one has about self-deception. It seems that it would be more straightforward to explain one's theory first, and give examples to support it afterwards. Yet if we do that, how shall we know what we are trying to explain? NST theories seem like a good place to start: they claim that the very concept of self-deception is not just paradoxical but inconsistent, therefore it can have no application. So NST theories need no examples, since they allege there can be no examples.

However, what NST theories give is a description of self-deception which they then try to show is inconsistent. This description is, in effect, a very generalised example, so general that it consists of nothing more than 'Someone is (literally) self-deceived'. Other theories can challenge NST theories by giving descriptions which (they claim) are consistent. These may be supported by alleged counter-examples - and <u>any</u> example which they offer will be a counter-example to the NST claim.

I have opted to discuss NST theories first, with their very general description of self-deception. Then I discuss Avoidance theories at the same level of generality. Before discussing Not-Know theories I offer three examples which, needless to say, are aligned to my views about self-deception.

The descriptions of the three examples are intended to exemplify three theories about self-deception. They are also intended to exemplify three ways in which we theorise about self-deception. Since I happen to be in broad agreement with the three theories, I also regard the three examples as correct characterisations of self-deception; but that is a claim for which I will have to argue subsequently. After offering my Process theory of self-deception, I offer a "case study": an extended example to show how the theory can be applied to particular instances of (alleged) self-deception, and a few sketchy examples drawn from other philosophical texts about self-deception. The extended example is drawn from the biographical and autobiographical book, Father And Son (Gosse, [1964]).

No-Such-Thing Theories

No-Such-Thing theories launch a triple attack upon other theories about self-deception.

1. They argue that the concept of self-deception is incoherent, and that the paradoxes of self-deception are symptoms of the concept's internal inconsistency.

2. They argue in detail against other theories, aiming to show that they cannot solve the paradoxes.

3. They argue that alleged examples of self-deception can be better explained by using some description other than 'self-deception'. Elster, for example, argues that alleged instances of self-deception can be more parsimoniously explained as instances of wishful thinking (Elster [1979] p149 - 152).

If a non-paradoxical account can be constructed, then items 1 and 2 listed above become irrelevant. Item 3 works only if the NST theorist's preferred theory really offers a better explanation. It must have at least as much power as the "self-deception" theory to yield true predictions and descriptions;

its "parsimony" must not be achieved by defining away parts of the data which are in need of explanation; it must not be paradoxical; and it must be distinct from self-deception. So, for example, if some instances of wishful thinking were cases of self-deception, Elster's "wishful thinking" hypothesis might turn out to actually exemplify "Such Thing" theories of self-deception, and not refute them at all.

The Paradoxes

The paradoxes are the strongest cards in the NST theorist's hand. Pears [1984] formulates the paradoxes as follows.

(1) If I have deceived myself that p then I believe that p but I also really know or believe or suspect that not-p: this combination seems impossible.

(2) If I know that the combination is impossible then I cannot intend to produce it in myself.

(3) If it is suggested that my fundamental belief is somehow screened from the rest of my thoughts and feelings, then the process becomes unintelligable since awareness of the belief is needed to motivate and guide the strategy.

(4) Perhaps then it is the whole plan which is screened, together with everything mental that it requires for its existence. But this merely shifts the paradox to a different point, at which it remains unresolved. If an internally coherent plan is impossible, it will not be made possible simply by my being unaware of it and not identifying myself with it.

I think that the paradoxes can be stated more strongly than this. First I shall suggest how Pears' versions of the paradoxes can be answered.

Regarding Pears' first point: the combination of believing that p and knowing or believing or suspecting that not-p is not impossible. It is quite commonplace for someone to have a set of beliefs which is inconsistent. So if there is anything paradoxical about having inconsistent beliefs, it is a paradox which is shared by many things other than self-deception. But why suppose that it is paradoxical? One may even know that one's set of beliefs is inconsistent, though without knowing which of the beliefs are inconsistent with each other. Pears suggests - in point (2) - that if I know that a combination of beliefs is impossible, then I cannot intend to produce it in myself. To which I add the rider that if I do not know that the combination is impossible then I can intend to produce it in myself.

Furthermore, someone might identify two beliefs which are inconsistent with each other, and have both beliefs while knowing that they are inconsistent. For the beliefs might be generated by circumstances beyond the person's control. Perhaps there are organic malfunctions in the person's brain which generate inconsistent sets of beliefs, pehaps the beliefs are caused by brainwashing, hypnosis, and so on. We regard it as undesirable to have mutually inconsistent beliefs, but the law of non-contradiction is a logical law not a psychological one: it is what we would like, not what we always get.

To make point (1) paradoxical, we need to add in more knowledge and more intention. For what I have suggested is that someone may be the unwilling victim of circumstances, and have inconsistent beliefs which he can do nothing about. In that case he is deceived, but not self-deceived: he is the victim of the deception but not its perpetrator.

These are the extra ingredients we need to restore the paradox:

- the self-deceiver believes that the two beliefs are inconsistent

- the self-deceiver intends to have both beliefs.

The paradox arises not when we have mutually inconsistent beliefs, but when we try to construct a strategy for acquiring the beliefs. For what is the intended goal? I can assert that p and assert that not-p, I can even assert that p and not-p - an outright contradiction - but asserting alone does not amount to believing. If I act on the supposition that p is true, then my action demonstrates that I do not believe that not-p is true. So what behaviour could manifest having mutually-inconsistent beliefs (rather than just asserting that one has such beliefs)? I suggest that if the beliefs cannot be demonstrated in outwardly observable behaviour, then they cannot be demonstrated in thinking, that variety of "inward" behaviour, either.

The answer to this paradox must lie in what the self-deceiver actually does: for that is what leads us to propose that he has mutually inconsistent beliefs. The self-deceiver does not manifest now one belief, now another - that is inconstant belief, not inconsistent belief. However we could argue that the inconstancy is the symptom of inconsistent beliefs, or of a paradoxical belief. Suppose, for example, that someone believes that the following sentence is true:

A. Sentence A is false.

If sentence A is true, then sentence A is false. If sentence A is false, then it is true ... and so on. This is a version of the "liar paradox". It is a very succinct instance of a paradox which could have much less succinct instances. The paradox can be created using two sentences:

B. Sentence C is true.

C. Sentence B is false.

Presumably similar paradoxes could be created using many more sentences. In that case one might work through very many sentences before the paradox received its "come-uppance" in the form of an outright contradiction. In working through all those sentences one might also be using them as hypotheses, one might be believing, acting, thinking, feeling, achieving goals by using them. Observing someone working through this sort of situation, we might argue that their behaviour manifests mutually inconsistent beliefs. For the set of beliefs is inconsistent, even though he has not worked through all the consequences in order to derive an outright contradiction.

This situation, though, looks like an instance of someone being mistaken, or confused. The situation is quite different to the proposed paradox of self-deception, in which someone starts out with the intention to have mutually inconsistent beliefs, and

therefore "knows what he is doing". He may indeed adopt the process of "working through consequences"; but throughout that process he knows what he is doing, so that instead of making genuine mistakes, and genuinely believing, he is merely pretending to believe. Merely pretending, is not genuine deception. So this attempt to characterise self-deception fails to avoid the paradoxes.

Our aim was to give a non-paradoxical description of a paradoxical set of beliefs. But in the case of self-deception, we are not the only ones who can describe the set of beliefs as paradoxical. The self-deceiver, if he knows what he is doing, can give the same description: in that case, he is not really deceived, but merely pretending.

Let us note here, for later reference, that by "working through the set of beliefs" the self-deceiver is not <u>at that time</u> able to describe the set of beliefs as paradoxical: for while he is "working through" them he is not yet in a position to describe them as paradoxical: that position is arrived at when he has worked through them sufficiently to arrive at a contradiction. So "working through" the beliefs is a way of <u>deferring</u> the knowledge of what he is doing. Deferred knowledge is a topic which will be addressed later on, in my discussion of Role Dissimulation theories.

In stating Pears' paradox (1) I have not mentioned the reason for suggesting that a self-deceiver has mutually inconsistent beliefs, namely that allegedly he uses a true belief to guide the strategy by which he acquires and sustains a false belief. This seems to me a very good description of how we use instrumental theories. We use a true belief to guide the construction of a false but useful simulation, in the sense given by the Shorter OED (Onions [1983]):

> simulation: the technique of imitating the behaviour of some situation or system (economic, military, mechanical, etc) by means of an analogous situation, model or apparatus, either to gain information more conveniently or to train personnel. (Onions [1983], p2660)

Constructing a simulation is one way of pretending. So this reinforces the suggestion that the self-deceiver is pretending to believe. To keep the paradox going we need to argue that the self-deceiver is "merely pretending", i.e. we need to argue that "mere pretence" never amounts to belief. I will suggest that some genuine beliefs (perhaps all of them) are pretences - in a sense of 'pretence' which I shall explain.

One may also use a false belief to guide the strategy by which one acquires a false belief: in this case one can presumably be self-deceived without "knowing the truth": one merely adds

another false belief to one's collection. I shall return to this point when I discuss Negligence theories of self-deception.

One may also use a false belief to guide the strategy by which one acquires a true belief. This has happened not infrequently in the history of enquiry. Presumably it is one way in which self-deception may cease. By using the false belief one arrives at the true belief which replaces it. If the process can go in one direction (from falsehood to truth), then it can go in the other direction too (from truth to falsehood). I shall return to this point later on. Here I am only setting down a marker for later discussion.

I have concentrated on what I take to be the strongest statement of paradox (1), in which the self-deceiver is supposed to have two beliefs which are inconsistent with each other.

However it is hard for me to understand why anyone would want to have inconsistent beliefs, and I have equal difficulty in trying to guess how that combination of beliefs would manifest itself in behaviour. Pears can help me out here since he claims that self-deception is (and must be) a manifestation of such a combination. But self-deception can be described and explained without postulating such a combination of beliefs, as

I shall later try to show. The point of self-deception is to replace one belief with another, not to sustain two beliefs which are inconsistent with each other. Therefore point (1) does not describe self-deception, and so it does not show that self-deception is paradoxical. Alleged paradox (2) collapses at the same time, since it depends upon paradox (1). Since self-deception does not require a combination of beliefs which are inconsistent with each other, there is no need to propose that some of the beliefs are "screened". Therefore paradox (3) collapses, and so does paradox (4).

Pears' versions of the paradoxes do not arise if we describe self-deception without appeal to inconsistent combinations of beliefs and "screens". However, Pears' formulations of the paradoxes are extremely effective weapons against theories of self-deception which do postulate the existence of "screened beliefs" etc.

If I have really deceived myself that not-p, then, contrary to point (1), I do <u>not</u> know or believe or suspect that p - not any more. The supposition that self-deception requires us to sustain a "fundamental" belief behind a "screen" seems to me unnecessary. Sustaining the belief behind a "screen" is perhaps the sort of thing someone might try to do because self-deception <u>failed</u>, because he failed to do away with an unwanted belief. The aim of self-deception (supposing that self-

deception has an aim) is not to produce an inconsistent combination of beliefs, but to replace one belief with another.

If point (3) shows that self-deception is paradoxical, then by similar reasoning we can show that shaving is paradoxical: for shaving is motivated by one's awareness of unwanted surplus hair. One must retain the hair in order to motivate the shaving. Shaving cannot remove the hair, for if it did, the motivation to shave would disappear.

I hope this "paradox" seems as unconvincing to you as Pears' third paradox of self-deception seems to me. Shaving is possible, and so is self-deception. The aim of shaving is not to produce an impossible combination of hairiness and hairlessness, but to replace hairiness with hairlessness. Likewise, the aim of self-deception is not to produce an impossible combination of beliefs, but to replace one belief with another.

One counter-objection is that shaving is not a fair analogy for self-deception. Unlike beards, beliefs can be re-constituted simply through the awareness of operating with them: it is as though shaving produced the hair it was supposed to remove.

The counter-objection assumes without any supporting argument that our way of replacing beliefs must "operate with" the

beliefs which are to be replaced. The production of the replacement belief may not use the belief to be replaced. I would expect that the self-deceiver would try to have nothing to do with that belief, and would certainly not make use of it in a way which builds it into the plan for a future without that belief.

Part of the argument is that the "fundamental belief" is needed to motivate self-deception. But the belief is not what motivates the self-deception; Dislike of the belief may motivate the self-deception: but that is a different thing.

Here is an illustration. Suppose I am socially inept, annoy people with my abrasive manner, and so on. I may dislike being like that, and try to do something about it. My strategy is unlikely to include any element of continuing to be abrasive, socially inept, and so on. If there is such an element, it is probably only to remind myself of how awful it is to be like that, thereby strengthening my resolve never to be like that again. Suppose I succeed in developing some tact and social graces: there is no reason to suppose that "deep down" I must still be socially inept or "really" still have an abrasive manner. The whole point is to do away with all that. The abrasive manner does not motivate the process: dislike of the abrasive manner does. Likewise, if the whole point is to do away with a belief, one is not likely to use the belief in

order to do away with it. One is not motivated by the belief, but by dislike of the belief.

If the aim of self-deception is to replace one belief (lets call it "the original belief") with another (lets call it "the replacement belief"), then the self-deceiver may know what the original belief is, know what the replacement belief is, and know that he intends to believe the replacement instead of the original. After successfully achieving the self-deception he may think, "thank goodness I do not believe that [the original belief] any more". Before and during the process he may think, "I refuse to believe it, and I shall find something better to believe".

The paradoxes can be more strongly stated as follows:

1. Suppose that someone, S, is self-deceived. S either knows what he is doing, or does not know. If he does not know, then he is merely mistaken. If he does know, then he is not deceived. Therefore the description of self-deception always collapses into a description of something different: being mistaken, pretending, deceiving others, or whatever it may be.

2. The self-deceiver needs to know what he is doing. If he did not, he could accidentally encounter evidence which

> would convince him or her of the truth which he was trying not to believe. But if the self-deceiver (let us call the self-deceiver "S", for brevity) knows what he is doing, then he is not deceived.

S may deceive others, but self-deception seems to be an impossible task. Not only must S know what he is doing, but it seems he must also know the very thing about which he is attempting to be deceived. For one needs to know it in order to avoid being confronted by the evidence for it: in order to ignore the evidence. But ignoring something is quite unlike being ignorant of it. One must be aware of it in order to ignore it. Otherwise, one may be confronted by it inadvertantly.

Ignoring something is a strategic movement of feigning ignorance, by putting oneself in a situation where ignorance would be possible (if one did not know already). We can detect the difference between someone who is ignorant of something, and someone who is ignoring something. For the behaviour of the latter is patterned around the thing he is ignoring, and this pattern exhibits the thing ignored as clearly as an archway exhibits the space left by the scaffold which once supported it: it is "glaringly absent". Someone who feigns ignorance is not ignorant, a feigned mistake is not a mistake, and someone who pretends to be deceived is not really self-

deceived. The behaviour which we describe as self-deception may be pretence, deception of other people, mistake, or ignorance. What it cannot be (according to this argument) is self-deception, literally understood. Consequently every attempt to characterise self-deception collapses into a description of something else.

Describing self-deception in this way invites paradox. It is rather like "the relaxation paradox", which goes like this: relaxation is impossible, for the attempt to describe it always turns out to be paradoxical. Consider someone who is trying to relax: the more he makes an effort to relax, the less relaxed he becomes. Making an effort is incompatible with relaxation. But he must make an effort in order to relax. For if he does not make an effort, he will never be able to achieve the goal, relaxation. Therefore the very idea of relaxation is incoherent. 'Relaxation' is always a misdescription, and the things we call relaxation are really something other than relaxation.

I hope that it is obvious that we do not need to make an effort to relax; and that relaxation does happen. If you want to relax, you had better not make an effort. Relaxation is achieved by not making an effort. Relaxation only seems paradoxical when we describe it in a peculiar way.

The "self-deception paradox" is similar. We do not need to know in order to be self-deceived; and self-deception does happen. If you want to be self-deceived, you had better not know. Self-deception is achieved by not knowing. Self-deception only seems paradoxical when we describe it in a peculiar way.

"The self-deceiver must know what he is doing" is like: "the person trying to relax must make an effort". Take away the assumption and there is no paradox. Instead there is a gap waiting to be filled with an explanation: the explanation we need is an account of what the person does instead of making an effort (in the case of relaxation), or instead of knowing (in the case of self-deception).

There is a variation upon the relaxation paradox. This variation too has a parallel in discussions of self-deception. We may be simply incredulous that someone is able to relax when he is placed in a very stressful situation: how can he be so relaxed when there is so much stress? The parallel in the case of self-deception is: how can he be self-deceived, when the deception is in conflict with such strong evidence? We are inclined to suppose that the evidence must coerce the self-deceiver into (or out of) believing something: the evidence must put an end to the deception, just as stress may put an end to relaxation. But someone may stay relaxed because he does

not regard the situation as being stressful; and someone may stay deceived because he does not regard "the evidence" as evidence. "The evidence" is what we call it; he may regard it differently.

The construction of the paradoxes depends upon keeping our distinctions sharp: feigning must be sharply distinct from being deceived, mistaken or ignorant. For otherwise we might describe a process by which someone slowly drifts between one state and another, say between feigning and being self-deceived, or between feigning and knowing; somewhere between the two there might be self-deception, not another sharply-characterised state but something which comes and goes by degrees. Also, we must sharply distinguish being mistaken from knowing. For otherwise we might describe (as in practice we often do) a process in which someone gradually comes to realise something (and he may be held responsible for not coming to realise it more swiftly), or gradually forgetting something, or gradually sinking into ignorance or error. If such a process occurs then one might occupy that twilight zone between knowqledge and ignorance, knowledge and error, for a long time, without ever emerging into one of the boundary states - knowledge, error, ignorance.

The paradoxes are constructed in a way that forces us to suppose that changes of belief nust always be discontinuities,

never allowing for a slow change between beliefs. For otherwise one might loiter between beliefs, not wholly convinced of any of them. Self-deception might then be described as a state of "half-belief", belief waxing and waning by degrees. So the paradoxes require us to suppose that there are no such continuities.

This is rather like Zeno's paradox of the arrow. The arrow can never move, because at any instant when it is in flight it must be at only one place. Since there is never an instant at which it moves, then if time consists of instants, the arrow cannot move. For there is never a an instant at which it is in motion. Or at least, the arrow cannot move continuously: it can only be in a succession of different states at different instants. The problem is that if this is true of self-deception, it will also be true of all beliefs: any change of belief will involve discontinuities; there will never be any gradual transitions from one belief to another. I think that while some changes of beliefs may be discontinuous, others are not.

I do not find the sharp distinctions intuitively plausible (which is not to say they cannot be correct). I am inclined to say we spend our entire lives in the "twilight zone" between knowledge, error and ignorance. What happens there is a process which we must later examine.

Despite what I have just said, I welcome the attempt to keep our distinctions sharp. For in this way, it becomes ever more obvious that self-deception cannot be either knowledge, or ignorance, or mistake, as "sharply defined". The instances of self-deception that we come across cannot then be explained as really being mistakes, pretences, or attempts to deceive other people. NST theories will then fail to show that these cases can be better explained without describing them as self-deception. For all other explanations will have been ruled out by the "sharp distinctions" which make self-deception unlike mistakes, pretences, etc.

However, the paradox can be reconstructed even if we replace the "discontinuities" hypothesis with a "continuities" hypothesis. Someone who is in the "twilight zone" between knowledge and error may have a combination of knowledge and error, but that does not compel us to suppose that therefore he is self-deceived. For to be self-deceived one must intend to be deceived, and if someone intends something then he knows what he intends: otherwise it is not intentional. If it is not intentional then it is merely a mistake.

Even if we propose such dubious psychological entities as "unconscious intentions", we do not escape the paradox. For being deceived by an unconscious intention is akin to being deceived by a parasite which has taken root in one's brain, or

having false beliefs because one has a brain lesion. Unless one can be held responsible for it - in rather a strong sense of "responsible" - one is not self-deceived but merely deceived. The fact that the cause of the deception is internal to the agent rather than external is not sufficient. One cannot be held responsible for something of which one is wholly ignorant. So, if we are held responsible for it, we must know what we are doing. If we know what we are doing then we are not really deceived - so not self-deceived either. The postulation of such things as "unconscious intentions" only appear to solve the paradoxes because they are not at all well-defined. As soon as they become well-defined the paradox reappears.

The construction of the paradoxes depends upon a sharp distinction between what is intentional and what is unintentional. In fact there are more distinctions we can draw. We can distinguish:

1. intending to do or to be something
2. intending not to do or to be something
3. not intending to do or to be something
4. not intending not to do or to be something.

We can distinguish further by reference to the "something" just mentioned. For example, we can expand 1. above:

1.1 intending to believe that p

1.2 intending to believe that not-p

1.3 intending to find out whether or not p is true

... and so on.

If self-deception is characterised as "not intending to do or be something" - namely, not intending to believe the truth - then self-deception is not intentional, and so the self-deceiver need not "know what he is doing". Although he lacks intention, he may still be responsible for being deceived. Negligence of something which one ought to do or be can be culpable. The self-deceiver neglects to believe the truth, and the truth is what one ought to believe, if one is able to do so (in some sense of 'ought'). Self-deception by negligence is not paradoxical. But it does not cover self-deceptions which are described by distinctions 1. and 2. above, i.e. where there is an intention and not just a lack of intention. If there are self-deceptions like that, they are not negligent. I shall argue that there are such self-deceptions, and that they are not paradoxical.

When we make our distinctions sharper, self-deception becomes more salient. For the sharper our distinctions are, the more obvious it becomes that self-deception is not knowledge, or ignorance, or unwitting error.

Are NST Theories Better Descriptions Than Such-Thing (ST) Theories?

Let us think about why we might want to describe something as an instance of self-deception. Suppose you have a theory about someone's aims. You observe them for a decade or two, and you notice that your theory accurately predicts their achievements. You also notice that their public statements of their aims do not match what they actually achieve. So your theory is inconsistent with their statements; and your theory works better than theirs.

There are a number of ways of explaining this. I label and itemise some of them below.

1. "Honest but inept": they stated their aims correctly, but failed to achieve those aims.

2 "Honest but inconstant": they stated their aims correctly, but the aims changed. However, your theory, yielding accurate predictions, suggests that there is a constant trend or pattern in their behaviour, which belies their statements.

3. "Lying": your theory is correct. Their statements are intended to deceive others.

4. "Speaking in code": you have mistranslated their statements. Correctly decoded, they would predict the person's behaviour as accurately as your theory does.

5. "Mistaken or confused": they are unwittingly mistaken or confused about their aims.

6. "Self-deceived": they are mistaken or confused or ignorant about their aims, in order to achieve those aims.

No doubt there are also other explanations which I have not listed.

There are ways of ruling out explanations too. For example, if the person shows no signs of disappointment and seems to thrive on his or her achievements, we may tend to think that he is not "honest but inept". For if he were, then he would consider them failures rather than achievements.

If he is "honest but inconstant", one would expect the statements of aims to alter, not just the behaviour. If he is "speaking in code", one can suggest a decoding. He may reject the proposed decoding. But if he cannot suggest a better one, then one may start to think that he is not merely speaking in code but is lying, mistaken, confused, or self-deceived. If the errors he makes consistently turn out to his advantage,

then one may start to rule out the "mistaken or confused" option: for this link to advantages seems too much of a coincidence.

This leaves the options, "lying", and "self-deceived". One can now forget about the person's own statements: the statements made by a self-deceiver or an other-deceiver are not a reliable guide to truth. Observing the person, we may ask which one is needed in order to achieve the aims described by our own theory. Does he need to deceive others, or himself, or both? In this way we may, by adding all sorts of circumstantial details into our description, narrow the possibilities down to a single explanation. The one I am interested in is, of course, "self-deception".

Such explanations may rarely proceed in so straightforward a manner. Someone may be self-deceived, lying, mistaken, confused, inconstant and inept, all at once. And so may we be, when we attempt to describe them. Sometimes, though, the situation may be relatively unambiguous, the description "self-deceived" being justified by a combination of:

- a theory which accurately predicts someone's behaviour

- a diagnosis of behaviour to discern emotional states (such as "not disappointed")

- a test of one's "translation" of a person's statements

- a description of the advantages accruing to someone because of their "mistakes"

- the strategy which would be needed to gain those benefits intentionally (e.g. is the strategy self-deception or other-deception)

- plus, no doubt, other factors which I have not considered here.

We can telescope all these factors into one item:

- your theory that another person is self-deceived enables you to accurately predict their behaviour, in contradiction to their own statements, and other theories do not enable you to predict as successfully as this one does.

In short, sometimes "self-deception" may be the only description that fits the situation. Since it works better than the alleged self-deceiver's own statements, you can argue that he ought to use your theory too. Not for his own benefit, but for ours: we are entitled not to be told something which can mislead us, and perhaps harm us.

Perhaps the alleged self-deceiver would be unable to achieve his goals if he used your thory. The factor which leads one to abandon self-deception may be a change of goals rather than an encounter with more evidence, gaining more information, losing an argument, etc.

When one's goals conflict with those of other people there may be a variety of outcomes. The self-deceiver may sacrifice the interests of others, or sacrifice his own interests. If we demand that the self-deceiver should adopt our theory, we are asking that he make a sacrifice in the interests of enabling us to have a more widely accepted way of making accurate predictions. Whether or not our demand is justified is another question. Our entitlement to not be harmed may not be absolute; and someone else's self-deception may not harm or mislead us at all.

An NST theorist might argue that "we truth-seekers" may have less than creditable motives for calling someone "self-deceived". For we may wish to blame people for their (genuine) mistakes: annoyed by the mistake, we may want to take it out on someone even though it was not their fault. By redescribing their mistake as "self-deception", we arm ourselves with a justification for blaming them. Another possibility is that we wish to exculpate someone for "mistakes" for which they deserve to be blamed (including our own "mistakes": "I must have been

deceiving myself" is more than halfway towards being an excuse).

One may wish to undermine the "self-deceiver's" arguments, drawing attention away from the arguments by finding fault with the person (which is irrelevant because a sound argument is sound regardless of the motives of the person who constructed it). We may even be defending our own self-deceptions by attacking the epistemic authority of other people ("he's in no position to judge: he's self-deceived").

The description, "self-deceived", may be the verbal trigger to provoke a response without needing to argue for it: a compressed argument which is never spelled out and which, if it were spelled out, might prove to be quite feeble.

The mere attribution of self-deception is insufficient: it has to be substantiated. But if it is substantiated, then the motives of the person who made the attribution, however discreditable they may be, do not detract from the truth of the description.

We may be able to construct a strong defence of the terminology of self-deception, if it works for prediction better than other ways of describing behaviour. And it certainly seems to be a disadvantage of NST theories that they force us to redescribe

every alleged self-deceiver as either mistaken or cynically lying to other people. NST theories force us to do so because they allege that the paradoxes of self-deception are insoluble: the very notion of self-deception is taken to be incoherent. NST theories are therefore vulnerable to the arrival of a non-paradoxical theory of self-deception. There is a variety of non-paradoxical theories available. Later on I shall argue that all the non-paradoxical theories are complementary (or can be made to be, by minor adjustments), and can be put together to form a super-theory.

I have not shown that 'self-deception' is ever a better description of some behaviour than all the other options. But the fact that 'self-deception' is part of our vocabulary suggests that we find some use for it, i.e. sometimes we think that self-deception is the best explanation of what someone is doing.

A House Divided: Schism Theories

> If a house be divided against itself, that house cannot stand.
>
> The Gospel According To Mark, 3:25

Sometimes one person deceives another. It seems unparadoxical. So if self-deception (intrapersonal deception) could be explained on the model of other-deception (interpersonal deception) then we could dispel the paradoxes associated with self-deception.

Schism theories in their simplest form consider the self-deceiver to be divided into two parts. One part is the deceiver, the other is the deceived. More complicated forms of schism theory propose that there are more than two elements within the self-deceiver.

Some schism theories explain self-deception as a process in which a person creates a schism within himself or herself in order to be deceived. Other schism theories treat self-deception as a way of exploiting a division which is already

there in the person, so that self-deception is more a matter of failing to integrate what was already divided.

Schism theories avoid the paradoxes by arguing that self-deception is not really reflexive.

Take the first sort of Schism theories first - those which claim that the self-deceiver creates a schism. So far as I have been able to find out, nobody knows how to create a schism within himself in order to be self-deceived. Supposing someone ever did achieve it, there is what I call the "Humpty Dumpty problem": once they are divided, nobody knows how to put the pieces back together again. So if schism explained self-deception, it would seem to be irreversible. Yet I believe that self-deception can cease, people can emerge from it.

When people cease to be self-deceived, they do not, so far as I am aware, report that a schism existed while they were self-deceived. But if there was a schism then one would expect them to be able to report their experiences from one side of the schism or the other, perhaps from both sides: for the schism is presumably no longer causing an intrapersonal communication problem.

Even if people did report the existence of a schism in their former self-deceptions, the report would not be irrefutable

evidence that any such schism existed. People who made such reports might be simply making their best attempt at self-explanation, based on the theories available to them. Since the theory available might be a schism theory, the report would be nothing more than an interpretation of what "must have happened" in the self-deception if the schism theory is true. The details of the report would be inferred from what the theory says must have been the case.

Creating a schism seems as problematic as deceiving oneself, and for much the same reason. Suppose that I could create such a schism. It would not help me to deceive myself because I would be on both sides of the schism. I might be able to arrange things so that the part of me on one side of the schism was able to deceive the part of me that was on the other side. But this arrangement would be very artificial; arranging for "part of me" to be "deceived" does not seem to help deceive me. So it might turn out that the easiest way to create a schism is ... to deceive oneself. However this order of doing things prevents us using the schism to explain the self-deception; but self-deception seems a lot easier than creating a schism.

So perhaps we are better off with the second sort of schism theories: those theories which claim that self-deception exploits a schism which is already there.

The schism itself is not the whole explanation of self-deception. If someone were deceived solely because of a schism which he had no control over, then there would be no self-deception: he would be the victim of the deception but not its perpetrator. He would be deceived but not a deceiver.

If schism alone were to explain self-deception, then there would also be a problem about real physical schisms within people who do not seem to be self-deceived. For example, a real physical schism is created by commissurotomy, an operation which divides someone's left brain from his or her right brain by cutting the connecting tissue between them. Wilkes [1978] discusses this in some detail.

If self-deception is schism, then people who have undergone this operation (sometimes called "split-brain patients") should be self-deceivers.

I do not believe that the commisurotomy operation creates self-deceivers. It can certainly cause intrapersonal communication problems - although these do not as a rule seem very apparent except under special experimental conditions. The sort of conditions which are created for the purposes of experiment are things like preventing the person seeing with his left eye what he can see with his right eye, and vice versa, i.e. ensuring that the "communication gap" between left brain and right brain

is not circumvented by using means of communication external to the person's brain. But these problems of intrapersonal communication seem like mistakes, not like attempted self-deception.

If a real physical schism is not what is intended when the self-deceiver is described as "divided", then there is a problem of knowing what sort of division is meant, and what is being divided from what. If we say that the self-deceiver's mind is divided, we have a problem about describing what a mind is, not to mention describing how it divides up.

Suppose we describe a set of mental functions which are divided. Are all the mental functions performed by one agent, or are they performed by two agents within a single body? If all the functions are performed by one person, in what sense could that person be described as deceived, and what is the point of postulating a division between the functions?

I am not suggesting at this stage that no satisfactory answer could be forthcoming: but it is a serious problem, which needs to be addressed.

To defend a Schism theory we would also need to be able to say which elements of the divided person went into which side of the schism.

Interpersonal deception is possible when one person (the deceiver) has privileged access to information which the other person (the deceived) lacks, or when the deceived grants that privilege to the deceiver.

Here are some examples of situations in which person A can deceive person B:

- A witnesses an event which B did not witness; B needs A's report in order to find out about the event

- A claims to be a witness to an event which B did not witness; B relies on A's report to find out about the event; A was not a witness to the event, but B does not know that A was not a witness - either B is not able to find out that A was not a witness, or B does not bother to find out (in which case there might be "contributory negligence" from B)

- Both A and B witness an event, but A claims to notice aspects of the event which B does not notice; A persuades B to accept A's account of what happened

- Both A and B witness an event, but A claims to have expertise which B lacks; A persuades B that A's "expert" account is a true account of what happened

- Both A and B are experts; A claims to have applied his expertise to generate an account of something; B does not apply expertise but accepts A's account

- A fabricates evidence and leaves it in a place where B will find it, taking care that B does not know that the evidence is fabricated: B finds the evidence and draws the false but inviting conclusions which A designed the "evidence" to "point to"

- A interferes with B; e.g. by getting B drunk, A prevents B thinking clearly; or by persuading B to depart from B's normal routine, A ensures that B does not encounter evidence which A wishes to keep hidden from B; or, when B is trying to decide what conclusion to draw about something, A intervenes, emphasising some arguments and playing down others, interrupting lines of thought which look set to lead to the truth while encouraging lines of thought which lead to false conclusions ... and so on.

In all these cases A has information which he witholds from B: the information may be what he really saw when he witnessed an event which he lies to A about, or it may be information about A himself, e.g. that he never really witnessed the event, or that it was A who fabricated evidence, and so on.

So, for example, you cannot deceive me into believing that Coventry City won the FA Cup this year if we are both present when the Cup Final is played and I can see with my own eyes that Coventry City lost; but if I retire to the bar in disgust at the way the match is going, and rely upon you to report further developments to me, then I have conceded to you privileged access which I have denied myself. You are then in a position to deceive me, should you wish to do so.

Suppose that you do not try to deceive me, but that I reject the information you give me. "What!" I roar, "City cannot be doing that badly - I don't believe you!" - but I do not make the effort to look through the window and check for myself. In this case the schism between deceiver and deceived explains the deception, not because one side is misleading the other but because one side distrusts the other in order to be misled. Instead of A misleading B, it is B who rejects A's information. In this case there seems little point in proposing a schism to explain the deception: for A performs no functional role in the explanation of how B came to be deceived. B does not need A's help in order to ignore the information. Instead of proposing that A supplies information which B rejects, we could simply say that B rejects, ignores, or ensures ignorance of the information.

There is one reason why A might not be entirely redundant in this scenario, namely that it can be much easier to reject hearsay evidence than the evidence of one'e own eyes. Even so, the hearsay evidence need not be supplied from within the self-deceiver, and I have not seen anyone else offering a schism theory which does not propose that one element is a deceiver and another element is deceived. Whereas this variant suggests that A is not a deceiver, but B wants to be deceived and exploits the schism between them to gain its objective.

If schism is to explain self-deception, we need to know what elements of the self-deceiver go on which side of the schism.

For example, suppose that self-deception about an event is like eyewitness A giving a false account to another element, B. A must have access to information which B lacks, e.g. information gained through the senses. We do say things like, "my eyes deceived me", but being "deceived by one's senses" can be distinguished from self-deception. It is one of the things someone might say in order to show that he is not self-deceived.

That is why it is important to attribute aims to the deceiving element within the self-deceiver. For if the deceiving element does not have aims, then we do not have a case of self-deception: the deceiving element functions in a "mechanical"

way, and the result is a mistake, not self-deception. The person is mistaken in the same way as someone who, due to poor eyesight, misjudges distances. If there is no aim to deceive, then there is no self-deception.

The deceived element must have aims too. For if all the person's aims reside in the deceiving element, then we would have to say that the person deceived an element within himself or herself which has no aims, which functions "mechanically".

We do sometimes say things like this. For example, a medical student who is about to witness a surgical operation for the first time may "steel his emotions": he makes an effort to view the operation in a mechanical, functional way, to focus attention away from the thought that the surgeon's knife is cutting into the flesh of another human being and the thought of how much that would hurt if the person were conscious, and so on. We could (stretching words considerably) say that the medical student is deceiving his emotions. But we would not be justified in saying that the medical student is deceiving himself.

So what goes into the seperate elements of the self-deceiver?

We could suppose that element A processes sensory information while element B makes decisions about what actions, if any, to

take. A influences B's decision-making by feeding false information to B, in order to secure a decision favoured by A.

Notice that quite a lot of extra assumptions have to be made for this sort of account to work. A is capable of much more than just processing sensory information: A also knows, or guesses, how B makes decisions: for otherwise A would not be able to decide what false information to feed to B in order to secure the desired decision. B must not process the sensory information directly itself, otherwise the deception will be discovered and defeated. And B must not know or guess that A is capable of practising such duplicity. For in that case the deception can only succeed with B's collaboration: which would make B not only deceived but also a deceiver. There would be no point in postulating a schism.

So A turns out to have a much greater role in the self-deceiver's psyche than B does. I described B as "making the decisions" as though B was a sort of powerful company executive and A was one of the company's information-gathering minions. But effectively A has taken control and is exercising executive power. In that case, we no longer seem to be justified in saying that the person is self-deceived; at most we could say he is self-deceiving: for executive power resides with A, A is the one which really makes the decisions, the one we should really identify as the major component of the person. If the

major component is the deciever and not the deceived, then we do not seem to have much of a case for saying the whole person is deceived. B is deceived, but the best way to describe this seems to be: "the person has made a decision to deceive one element within himself or herself: that element is B." B is deceived, but the person is not.

Notice the extreme artificiality of the terminology we are forced to use in describing the schism. We are obliged to think of elements A and B as two agents, both having their own goals and wishes.

Lets try out the idea that there are two agents within the one person, each agent being "complete" in the sense of having all the information-processing faculties and aims that a whole person might have.

"Eye-witness" style deception is obviously impossible, since both agent A and agent B share the person's senses. However A might use sensory information which B does not use. And "expert" style deception can occur if A is an expert and B is not, or if both are experts but B does not apply its expertise and relies instead upon A.

If A and B also share whatever activity the person undertakes then there is no chance of A planting evidence while B is not there.

So let's try out the idea offered by Pears [1984]: that A is an agent which is within B. B observes its environment through the senses, and acts within that environment; whereas A does not have direct contact with that environment. A's environment is B: A is encapsulated within B. A knows about B and acts within B. A's aims are different to B's. For example, B wants to know the truth, but A wants B to be happy; and in the case of the self-deceiver, A knows that the truth is incompatible with B's being happy. A therefore acts upon B to change B's beliefs. Because A has this power to alter B's beliefs, B is unaware that A exists and that A is busy altering B's beliefs. A deceives B.

I wonder if we can call this an instance of self-deception. If B had a brain tumour which caused B to have false beliefs, we would not call B self-deceived. The consequences of A's activities are perhaps like those of the brain tumour. However there is a difference: unlike the brain tumour, A has aims; the alterations to the beliefs are not random, they form a pattern, with the design of keeping B happy.

The situation reminds me of a common variety of science fiction plot, in which extraterrestrial creatures invade from outer space and take up residence in people's brains, imposing their own aims upon the people. B is certainly imposed upon by A ("for its own good", no doubt). A is one element of the whole person. A is also the element which is in charge: B is its helpless dupe. But A has no direct access to B's environment, including other people.

The role played by A in the whole person's psyche seems rather like the role of a parasite which has intentions. Someone who was controlled and deceived by a parasite would not be described as self-deceived; but in this case the "parasite" grows from within; a brain tumour also grows from within, but we do not call someone self-deceived if their beliefs are being warped by a brain tumour.

B cannot detect A. But perhaps other people can detect the effects of A's activity. There is a pattern in B's beliefs which they can detect, and they can communicate their recognition of the pattern to B (remember that they cannot talk directly to element A).

B's beliefs are now being tugged by two conflicting forces: A, which aims to remain concealed, and other people, who are pointing out the effects of A's actions So it seems possible

that B can, momentarily at least, become indirectly aware of A. B can then, through perception of the whole person's behaviour, indirectly trace A's activity, and take steps to counteract it. This is assuming that A can only influence B's beliefs, and does not wholly control them.

If A wholly controls B's beliefs, then the theory becomes vulnerable to the objection that B is "mechanical", that the whole person can be identified with A, and therefore that the person is deceiving an element within himself or herself but is not self-deceived.

Pears has breathed new life into Schism theories with this suggestion. However, I wonder how much of a role the "schism" actually plays in the "inner agent" variety of schism theories.

The inner agent, A, exists within B. A is able to intervene in its environment because it is part of that environment, not divided off from it. So A is a part of B, but a part which is detectable only indirectly, through its effects. Are the effects sufficient justification to warrant our positing that A exists, and is divided from B by a schism?

A is an agent with aims. Its effects are the pattern in B's beliefs. B also has aims, e.g. to have true beliefs. Instead

of proposing that any such thing as A exists, we could say something like this:

- the whole person (the alleged self-deceiver) has a mixture of aims, some of which are in conflict: he wishes to be happy, and he wishes to have true beliefs

- sometimes having true beliefs is incompatible with being happy

- he pursues the aim of having true beliefs through a conscious process of assessing evidence, and so on

- he consciously pursues the aim of being happy, where possible, by acting to alter the environment

- sometimes, though, there is nothing to be done by way of altering the environment which could avoid unhappiness: e.g. the things which cause the unhappiness are in the past, and the past cannot be altered

- yet he wishes the past were different

- there is a mental process by which wishes can simply be transformed into beliefs

- this process can occur spontaneously, without the need to perform any special actions or do any special conscious thinking

- the process happens whenever the person does nothing to prevent it

- therefore there is no need to propose the existence of an unobservable entity, element A, which is an agent internal to element B; because there is a tendency for wishes to become beliefs, there is no need for an agent to organise and carry out this process: it happens spontaneously.

The seperation between the individuals involved in interpersonal deception does not strike me as an important part of the explanation of how the deception is possible: if you and I were entirely unable to contact each other, for example, then neither one of us could deceive the other, no matter how hard we tried. So the most important part of the explanation in Schism theories seems to me to be the explanation of how the schism is overcome. And that raises the question of whether the schism itself has any explanatory value at all.

Yet the ability to withold information, particularly about one's intentions, does seem to be an important factor if someone is to succeed in deceiving other people. If you could

clearly recognise my intention to deceive you, for instance, then you would be forewarned and forearmed against my attempt. This does not mean that in such circumstances one person could not deceive another. It is quite possible to deceive people after having informed them that one is going to deceive them. They may not take the warning seriously, so they may be deceived despite it. But if you were aware of my intention to deceive at the instant when I was engaged in it, my attempt would be unlikely to succeed (though it might: if you were a self-deceiver it might suit you to collaborate in being deceived by me, for example).

So how does one person deceive another? The deception can be achieved through their behaviour, through arrangements of physical objects, by fabricating evidence to "point" to a false conclusion. But - let us ask an apparently naive question - if the deception is powerful enough to convince the victim of the deception, why does it not convince the deceiver too?

Suppose my behaviour deceives you but it does not deceive me. Why not? Perhaps I have information which my behaviour does not disclose to you.

We can give this case a "behaviourist" twist. We can argue that the way I find out about myself is exactly the same way that you find out about me, i.e. through my behaviour, as

directly observed by you or as reported by third parties. Then if my behaviour deceives you it will deceive me too. If I am not deceived, then (dropping our temporary conversion to Behaviourism) maybe it is because I have private access to my own thoughts which you are denied. But my thoughts may be no more transparent than my behaviour. They may not be any more "opaque" than my behaviour either. My body language may express my thoughts as clearly as if, every time I think, a big bubble appears over my head expressing my thoughts in words and pictures. I might still be able to deceive you, by deceiving myself.

Unless I have a "transparent" understanding of my own thoughts, unless I am incapable of being mistaken about the reasons for which some thoughts pass through my mind, then I can deceive myself by the way I think just as easily as I can deceive you by the way I act - perhaps even more easily. For whereas you have your own independent thought processes with which to assess my behaviour, I do not have a thought process independent of my thoughts. In that case self-deception might be explained not by divisions within the self-deceiver, but by the self-deceiver being too well-integrated. Total integration would mean the lack of an independent thought process with which to criticise what may seem to other people a clear case of self-deception. I shall pursue the claim that the self-deceiver is "too well-integrated" in a subsequent chapter.

Plato's Bestiary

> Imagine a very complicated, many-headed sort of beast, with heads of wild and tame animals all around it, which it can produce and change at will ... add two other sorts of creature, one a lion, the other a man. And let the many-headed creature be by far the largest, and the lion the next largest ... then put the three together and combine them into a single creature ... then give the whole the external appearance of one of the three, the man, so that to eyes unable to see anything beneath the outer shell it looks like a single creature, a man. (Plato [1974], p416, 588 c - e)

Plato portrays the human individual as consisting of a collection of animals. "Plato's bestiary" is my name for this portrayal. He takes the most characteristic, distinguishing element of a human being to be "reason", and so this element is represented in Plato's bestiary by a man.

Plato's view is that when the beastly elements in the individual overwhelm the most human element, the result is injustice: disharmony, disorder and deception. The individual can take steps to prevent injustice within himself, or he can

allow it to happen. If he allows it, the result is a loss of freedom; the loss of freedom is related to, but not identical to, a loss of responsibility. The individual loses freedom because what is most characteristically human about him becomes enslaved to what is bestial. Injustice may be the result of his upbringing, the kind of society in which he lives, or his own choice: there can be degrees of responsibility.

"Degrees of responsibility", in my reading of the text, equates to the degree to which injustice is self-inflicted or inflicted by others. If it is self-inflicted, the individual is also self-deceived: if it is inflicted by others, then the individual is deceived but not self-deceived. But unless the individual is a helpless victim and not a responsible agent at all, there will always be some degree of self-deception.

In the previous chapter I suggested that, despite what Schism theories say, self-deception may be better explained by proposing just the opposite: that the self-deceiver is too well-integrated. Plato offers us such a theory. The theory says that there is - and ought to be - a division of functions within a person. Self-deception arises when the person allows one function to encroach upon the activities which properly belong to another function. Plato suggests this kind of functional division: that people are internally divided into three functional parts: reason, ambition (there is no single

word which exactly translates the word used in the text, but 'ambition' is my preferred option) and thirdly the appetites. Plato claims that society is divided into three classes: people can be assigned to their class on the basis of which of the three parts is dominant in their character.

> The just man will not allow the three elements which make up his inward self to trespass upon each other's functions or interfere with each other but, by keeping all three in tune,like the notes of a scale ... will in the truest sense set his house to rights, attain self-mastery and order, and live on good terms with himself. When he has bound these elements into a disciplined and harmonious whole and so become fully one instead of many, he will be ready for action of any kind (Plato [1974], p221, 443d).

Each part has characteristic aims. Reason, for example, aims for truth. The appetites are physical and instinctive cravings. Ambition aims for honour: ambition has characteristics such as pugnacity, enterprise and indignation, which are often found in conflict with unthinking impulse. Ambition aims to win battle honours in war, to win arguments in disputation, and so on.

Plato's view is that Reason should command. But it does not always do so. Sometimes, for example, the appetites have the

upper hand. The appetites, in Plato's view, are a disorderly lot and not at all amenable to reason. They can be reined in or given their head; some of them are "unnecessary" and can be "killed". Some of them are necessary (i.e. you cannot kill them off) but you can subdue them by starving them. That is about all that can be done to control the appetites. If you indulge your appetites, however, they will grow stronger and plenty more of them will spring up. The appetites are in conflict with each other as well as with reason and ambition. Someone aiming to be "just", in the sense of "justice" quoted above, will aim to prune down the appetites until they form a collection which can be satisfied with the minimum amount of conflict.

When reason, whose proper role is the pursuit of truth, is subservient to the appetites, it will be engaged in such activities as calculating the way to maximise rewards for the appetites. Or ambition may be dominant, so that Reason is engaged in devising strategems for winning honour, whether it be in battles or verbal disputes, in commerce or elsewhere. For example, the oligarchic character,

> elevates the element of desire and profit-seeking to the throne [where it governs like] an oriental despot with tiara, chain and sword ... while reason and ambition squat in servitude at its feet, reason [is] forbidden to make

> any calculation or inquiry but how to make more money...
> Plato [1974] p370, 553c - d

When reason is subservient to another part of the person, then, considerations of truth are neglected; because reason is not performing its proper function, the person acquires false beliefs. The role of reason (to command) has been usurped. Someone who allows this to happen, and perhaps revels in it, is deceiving himself.

There are some problems in this account. Plato portrays the role of reason as being something like the role of a charioteer, directing and reining in the horses when necessary in order to reach his objectives. But how does reason do so? reason has the power to reason well: but reason cannot very well appeal to the appetites, for example, to be reasonable; by definition, they are not.

Reason can hardly overwhelm the appetites by force, since force is one of their dominant characteristics. Fortunately there is the third element, ambition, which under the tutelage of reason can strive for excellence and which may also be strong enough to rein in the unruly appetites. Ambition, then, is available as a mediator between reason and the appetites. But how does reason win the sympathy of ambition? Perhaps ambition is able to appreciate the arguments put forward by reason; but if

ambition is amenable to reason, that seems to indicate that ambition can distinguish sound arguments from unsound arguments. Perhaps ambition, then, is capable of reasoning on its own account, while the element which Plato calls reason might (against the spirit of Plato's text, to be sure) be described as an appetite for knowledge. It looks as though the three elements in the soul, which are supposed to be wholly distinct from each other, must share each other's characteristics to some degree.

How is reason to prevail? Plato uses the tripartite soul to give a metaphor for the tripartite state. He tells us: "our rulers will have to employ a good deal of fiction and deceit for the benefit of their subjects" (Plato [1974],450 c); spoken falsehood can be used "as a kind of preventive medicine against our enemies, or when anyone we call our friend tries to do something wrong from madness or folly" (Plato [1974], 382c). So, taking the tripartite state to be a metaphor for the tripartite soul, I take Plato's view to be that reason is to prevail by using fiction and deceit to rule over the lion of ambition and the many-headed dragon of the appetites.

I find this fascinating. Schism theories claim that self-deception is due to one element within the self-deceiver misleading another element. Now we find Plato claiming that one element, reason, must "bewitch" or deceive the other

elements in order to prevent self-deception: self-deception arising when the other elements prevail over reason. The other elements, being unable to reason, will be all the more vulnerable to such trickery, of course.

Yet this picture poses problems. The appetites, at least, are presented as being not just poor at reasoning, but wholly unreasoning. Therefore an attempt by reason to bewitch or deceive the appetites would be like someone trying to deceive the force of gravity, or the east wind, or an amoeba: they just are not the sorts of things which can be deceived. You cannot make the appetites think something which is false, for they are not thinking things. So how does reason prevail?

Lets try another tack.

> Suppose a man was in charge of a large and powerful animal, and made a study of its moods and wants; he would learn when to approach and handle it, when and why it was especially savage or gentle, what the different noises it made meant, and what tone of voice to use to sooth or annoy it. All this he might learn by long experience and familiarity, and then call it a science ... but he would not really know which of the creatures tastes and desires was admirable or shameful, good or bad, right or wrong; he would simply use the terms on the basis of its

> reactions, calling what pleased it good, what annoyed it bad. He would have no rational account to give of them, but would call the inevitable demands of the animal's nature right and admirable, remaining quite blind to the real nature of and difference between inevitability and goodness, and quite unable to tell anyone else what it was. (Plato [1974], 493a - c).

This seems like a way in which reason might prevail over the appetites: reason might prevail over the appetites in the way a keeper prevails over an animal. But Plato immediately disabuses us of this notion, comparing the keeper to "the man who thinks that the knowledge of the passions and pleasures of the mass of the common people is a science": "He is going out of his way to make the public his master and to subject himself to the fatal necessity of producing only what it approves." (Plato [1974],493d). If reason acts as a keeper of the appetites then it will end up subservient to them.

Plato offers a gentler image of the tripartite soul too, an image which is more helpful in answering our question, "how does reason prevail?" Reason is compared to a shepherd, ambition is compared to a sheep-dog, and the appetites are compared to sheep.

> It would be the most dreadful disgrace for a shepherd to keep sheep-dogs so badly bred and trained, that disobedience or hunger or some bad trait or other led them to worry the sheep and behave more like wolves than dogs ... we must therefore take every possible precaution to prevent our auxiliaries treating our citizens like that (Plato [1974], 416 a - b).

The "auxiliaries" in Plato's tripartite state are the equivalent of "ambition" in the tripartite soul.

Sheep are not easily trained, and the shepherd does not attempt to train them. The shepherd trains the sheepdog, which is apt to be trained; he uses the sheepdog to control the sheep. Reason can train ambition by coaxing and shaming it, habituating it to seek excellence.

Plato's account makes a lot of good points. One of the drawbacks is that it appears to require that there be three seperate entities or faculties, and the idea of seperate faculties is for good reasons not a popular one today. But this may be a misreading on my part: the words used by Plato can be translated as describing three seperate "forms". Anyway, the idea of the reason-ambition-appetite relation can be translated from the terminology of faculties into that of motivation / function. We could say: here are a set of

motives, some of which can be modified by education, some of which can be modified by training, some of which cannot be modified. Intelligent responses are typically modifiable, not rigid. The rational motives - such as the desire for truth - can be swamped and overwhelmed by other motives, such as those of appetite and ambition. Self-deception can be the result when physical and instinctive cravings or the ambition for honour and "success" gain greater priority than the rational desire for truth. Since we are not suggesting that there are distinct faculties, we do not need to propose that the different faculties overwhelm or trick each other.

Plato's distinctions between the different roles which may be played by reason are preserved in the English language by a variety of expressions. Consider these expressions, which range from derogatory to approving: 'crafty', 'calculating', 'skilful', 'clever', 'knowing', 'knowledgeable', 'wise'.

'Crafty' is reason in the service of Plato's class of craftsmen/artisans. 'Calculating' is reason serving the artisans or the "ambitious" guardians, the military guardians. 'Wise' is reason serving the characteristic aims of reason; and so on.

In Plato's account of the tripartite soul we find one element deceiving others - but this situation is not the one which is

considered to be self-deception: Plato's theory is not a Schism theory of self-deception. Self-deception in Plato's theory is a consequence of "injustice": when the three elements in the soul are not harmonised, and all three are thwarted in the pursuit of their proper aims. Injustice leads to self-deception and the division between the three parts (or three motives) is what makes the disharmony and disorder of injustice possible. But the division does not explain the self-deception. Nor does the self-deception explain the division. The deception arises as a result of the loss of freedom, it does not precede it; for, Plato maintains, "no man wants to be deceived in the most important part of him and about the most important things; that is when he is most terrified of falsehood" (Plato [1974], 382a). Self-deception occurs because reason is interfered with by the appetites and by ambition, the "spirited" part of the soul.

I have described Plato's account as a "tripartite theory of the soul", and Plato's text does often read as though there were three elements battling it out within the soul. We may wonder what place is left in this picture for the person to make a choice or have a preference as to the outcome of the battle. It seems as though only the three elements make choices or have preferences; but in other parts of the text we find that these three elements are supplemented by a fourth, as in this passage: "a man of sound and disciplined character, before he

goes to sleep, has wakened his reason and given it its fill of intellectual argument and enquiry; his desires he has neither starved nor indulged, so that they sink to rest ... the third, spirited, part of him he calms and keeps from quarrels so that he sleeps with an untroubled temper" (Plato [1974], 571d - 572a). Here there is a fourth element at work: the one which wakens reason, neither starves nor indulges the appetites, and calms the spirited part of the man. Faculty psychologists after the time of Plato have given this fourth element a name: the will. It is the will which is the executive part of the soul, which integrates (or divides) reason, ambition and appetites or which, in some accounts, replaces them.

I should add that I do not believe in these seperate faculties. We can gain the same benefits from Plato's account if we recast it in this way: instead of Plato's "three parts of the soul" we focus on his "three sets of motives ... knowledge, success or gain" (Plato [1974], 581b - c). We can divide our time and resource in pursuit of any one of these, any two, or all three. Usually our motives are mixed; sometimes our combined motives are incompatible with each other because it is not possible to achieve all of the aims together. Sometimes we can choose how we spend our time and resource: it is up to us. Sometimes we may not have the choice: perhaps we can be helplessly "overwhelmed" by our "appetites", for instance. At other times we may "give in to" physical cravings or by a desire to be

honoured although we are not "overwhelmed" by them, and may have second-order desires which conflict with them (e.g. an example of a second order desire would be the desire of a heroin addict not to crave heroin). Then we may "not face up to" what we have done: if we devoted time and effort to our motive to achieve success, we might feel ashamed of what we did; or we might feel elated. What I am getting at is that these are all situations in which we can try, put in effort, achieve, fail - they are all situations in which it makes sense to say that we are (to some degree) responsible for what happened, that we did something at will: these situations do not just befall us: we contribute to them, develop them, bring them about, and influence their outcomes. One of their outcomes is that we acquire, sustain, alter or shed beliefs. And so we are responsible for our beliefs; therefore when we are held responsible for having false beliefs we may be called self-deceivers.

One of the claims made about self-deception is that "the self-deceiver must know what he is up to". Very possibly someone may know what he is up to, if he is spending time and resource on trying to find out what he is up to - i.e. if he is pursuing the aims which, according to Plato, are characteristic of "reason". But if he is not engaged in that branch of enquiry, there is no reason why he should know: indeed it is impossible that he could know, because there is no process going on which

would generate that knowledge. Perhaps there ought to be such a process going on: and perhaps we hold the person responsible for not knowing because we blame them for a sort of "epistemic negligence". Nonetheless, the brute fact is that no such process occurs. The self-deceiver is up to something and is getting false beliefs as a result, but whatever he is up to is not what Plato claims is a characteristic aim of reason, namely the pursuit of knowledge.

So here is a situation where someone acquires false beliefs and we hold them responsible for doing so and where, incredulity of bystanders notwithstanding, he does not know that it is happening. He is self-deceived. But this self-deception is not paradoxical. There is no suggestion that the self-deceiver "really knows that what he believes is false", only that he could know, if he put his mind (and other resources) to it. He does not acquire the beliefs as a result of seeking truth: if he did then he would be merely mistaken, (a failed truth-seeker), not a self-deceiver (who is not seeking the truth at all). He has the beliefs because it suits him to have them.

I need to answer a possible objection here. The objection is this: it is misleading to call these beliefs, for to believe something is to believe that it is true. But the self-deceiver does not acquire the "beliefs" as a result of seeking for truth, therefore he cannot believe they are true.

My reply is this: truth-seeking is not the only way to acquire beliefs - why suppose that it is? The supposition certainly makes self-deception paradoxical. It also makes the paradox impossible to solve. For if the self-deceiver is responsible for having false beliefs, and has those false beliefs for a motive (because those beliefs suit him), and truth-seeking is the only way of acquiring beliefs, then there is no way in which self-deception could be achieved. If he does not seek truths then he will not acquire any belief and so will not acquire the false belief; if he does seek truths then either he will acquire true beliefs or he will acquire false beliefs but only by mistake - and a mistake is not self-deception; nor can he make a "deliberate mistake" in seeking truth: for then he would not genuinely be seeking truth, and so whatever he acquired would not be a genuine belief: furthermore since he is engaged in truth-seeking, he would know that whatever he acquired was a sham belief, not the genuine article.

Plato's Bestiary contains many themes which deserve more development, such as: (i) self-deception through neglecting to do something one ought to have done; (ii) mixed motives, with motives being linked to beliefs; (iii) mental processes, (of which truth-seeking is only one), generating beliefs; (iv) responsibility for one's beliefs, and of being able to do something to acquire, sustain, alter or shed one's beliefs. These themes are developed in the next few chapters.

"Too Well Integrated"

Lets indulge in a rather fanciful example. Suppose that my eye is glued to a telescope, so that to look at things I have to look through the telescope. I can alter the focus of the telescope, and I can move it about so as to look at different things; but I cannot take my eye away from the telescope. Suppose also that I am unable to open the other eye. So I have no independent means of visually checking the accuracy of the telescope.

In epistemology there is a tradition of questioning the veracity of the senses. The scenario presented is that I am intimately connected to my senses, as though "glued" to them. I cannot see independently of my eyes - just as, in the above example, I could not see independently of the telescope. My other senses may provide independent means to check the accuracy of vision; but I have no means independent of all the senses by which I may check the veracity of the senses. So perhaps my senses are not sources of information, but only serve to deceive me.

The sceptical argument about the senses can be extended. For the veracity of my thinking may also be put in question since,

after all, I cannot think independently of my mind. Just as I can focus and move the telescope, so I can focus attention on some things, and direct attention away from other things. But it seems that I have no independent means of checking what I think. I can only think the things which I have the aptitude to think, and my mind may distort information.

We may also doubt the veracity of people's first-personal reports of what they think. Perhaps I can give authoritative reports of what I think; but my reports, though authoritative, may be false. If we think by using a language, then we are vulnerable to its faults. It may be a distorting medium, a generator of falsehoods.

If I am "well integrated" then I will lack an independent means to check the truth of what I believe. Having no means of checking that is external to the means we have, the most we can achieve is internal consistency - cosistency between the media at our disposal - thinking, the senses, language. Within this limit, though, we have considerable room for manouvre. We do not know the limits of what we are able to think. Our theories are always vulnerable to the arrival of new and better theories which may replace them.

By trying out a different theory, we gain another means of checking the theories we have. If the theory is better (e.g.

it provides more internal consistency) then we may swop to using that theory permanently - or until it is challenged by yet another new and better theory.

We are not compelled to try out new theories. We may confine ourselves to the theories we know and love. We will not then suffer the "schism", the inner conflict, of having two ways of thinking, neither of which is consistent with the other.

We do not know what a new theory has to offer until we try it, if only to the extent of "entertaining the idea" that it may be true. If we do not try it, then we do not suffer any schism: we remain "well integrated". But we may be denying ourselves benefits which the new theory could give us. In particular, we may be denying ourselves truths - deceiving ourselves.

The scenario which I am constructing here is one in which self-deception is due not to schism but to being "too well integrated". The example I gave, namely of refusing to entertain a new theory, is a conservative variety of self-deception: bigotry. The bigot's motto might be:

> I do not know, and I do not care to know.

There is a sense in which the bigot can accept this motto, and a sense in which he cannot. The former sense is:

> I do not know about this theory, and I do not care to know about it.

The latter sense is:

> I do not know because I do not use the theory: I deny myself the means to knowledge.

The bigot cannot accept this because he would deny that the theory is a means to knowledge.

A less conservative self-deceiver might use the theory in order to put his or her old theory in doubt: now he has two ways of thinking, inconsistent with each other, both susceptible to sceptical attack. Since apparently neither theory is better than the other, he has the benefit of being able to pick and choose between them, using now one, now the other, at will.

This looks like a schism, but he is not divided: he is playing different roles in turn, "vacillating", and he does this as a coherent strategy. I think that this is a form of self-deception. He takes care not to create a decisive confrontation between the theories, so that he preserves the freedom to pick and choose. He neglects to do what scientists often try to do: find an experiment, a test case, which will decisively refute one or other of the contending theories.

There is another kind of opportunity for self-deception, too. He may "leap" into the new theory, that is, abandon the old theory and start using the new theory forever more. He does not bring about a decisive confrontation between the two theories. He is now a "bigot" who believes the new theory. He knows all about the old theory, having used it in the recent past. But he does not use it now.

Like someone who possesses a telescope but does not use it, he is deprived of the information it could have provided. This seems to me an instance of self-deception. Even if the new belief is true, it was not obtained in a truth-regarding way. The means used to arrive at the new theory were deceptive.

In a later chapter I discuss what "leaping to conclusions" comprises. For now I want to follow up the characterisation of the "bigot" who, instead of being "open-minded", has a "closed" mind.

He may be prepared to discuss other points of view, but always from the unchanging perspective he already has. He never adopts another point of view, not even provisionally, as an experiment. He can never be persuaded of what he never thinks, and since he never tries out a new theory, he never thinks it.

Adopting another point of view (even provisionally) requires us to make some suppositions. "But why," the bigot may ask, "should I suppose something which I do not believe? Why should I pretend that something is true when I do not believe that it is? Isn't that the way one slips into self-deception?"

Why pretend? The answer, as I shall argue in later chapters, is that without such "pretences" we cannot acquire beliefs at all. The bigot has forgotten (or chooses not to recall or to hypothesise) that the beliefs he has were to begin with constructed by making suppositions, "leaping to conclusions", and using them as if they were beliefs. If he has made suppositions in the past - in order to try them out, in order to acquire beliefs - what is the justification for refusing to make suppositions again? One reason for refusing to do so is that he is happy with the beliefs he has; but this reason, which explains the refusal, does not, in my view, justify it.

Perhaps he believes despite strong evidence against the beliefs. Consider how he might assess the evidence (I am assuming that he does not simply ignore it). He constructs a logical derivation taking the beliefs and the evidence as assumptions. He derives a contradiction. By reductio ad absurdem he is entitled to reject one or more of the assumptions, and he does: he rejects the evidence on the basis of the beliefs.

The evidence may be obviously true (to other people). What is obvious to him is that it is inconsistent with his beliefs. So (to him) the evidence is obviously false. He may or may not construct an explanation of why the evidence is false. If he does, the explanation will be consistent with the beliefs.

Hence the bigot's beliefs are unshakeable. For he never engages in the activities which could shake them. This is not a case of failing to indulge in critical enquiry. He does so - after a fashion. He constructs logical arguments in which the evidence is refuted and the beliefs are sustained. He does not engage in critical activity in a way which could put those beliefs at risk.

However there are other ways in which the beliefs may be shaken. They may fail to work well to achieve what the bigot wants. A crisis then ensues in which the bigot's strong attachment to the beliefs is jeopardised by an equally strong attachment to the things he wants to achieve. In these circumstances the bigot's self-deception may be destroyed because the motivation for it is destroyed.

I have implied that there is an activity which can "shake" beliefs. I also hinted that this activity involves an imaginative leap: we have to make suppositions: suppositions

which need not be counterfactual, but which may contradict our beliefs.

Making an imaginative leap is something we can do at will - or fail to do, at will. So the construction of beliefs can be a directly voluntary activity, and not just "indirectly voluntary" as some people (e.g. Audi [1982]) have suggested.

This does not mean that we are totally free to "leap". There may be limits to what we are capable of imagining. There may be horrible threats as to what may happen to us if we do leap - or if we refuse to leap. We may perceive that the leap tends towards beliefs which would cause us pain or grief. If we leap to a false belief, and act upon it, it could kill us - though not as invariably as epistemologists sometimes suggest.

Preserving one's integrity - the unity within oneself - may be less valuable than it is often claimed to be. Intellectual honesty may lead to deep inner conflicts, the disintegration of the much-vaunted "well-integrated" personality. The result may be a more genuine understanding of oneself and the world. A failure to avow the activities in which one is engaged may be a phase in a movement towards authenticity and not, as Fingarette [1969] and others have suggested, self-deception.

To hesitate in the divided, conflict-ridden phase of this move need not be self-deceptive. One may be genuinely unsure of what to do - perhaps of what one should do. It may take time to summon up the courage to "leap": one's sense of one's identity may be at risk. The duration of the phase, and its direction (towards a leap or away from it) may be important factors to consider, when we try to assess what is going on.

I do not wish to give the impression that self-deception consists of a failure to make imaginative leaps. Someone may hop from one theory to another in order to avoid a persistent suspicion which pursues them, or perhaps in order to evade a belief which a combination of circumstances and motives will make inescapable if he ever settles down. For just this reason, the refusal to leap may be the "authentic" thing to do: to "dig in" and wait for an understanding to develop. Our strategies for thinking are complicated and adaptable. Each case has to be assessed on its own merits. There is no limit in principle to what the assessment needs to take into account. In this respect a discussion of our cognitive activities is like the moral evaluation of actions. We cannot rule out the possibility that the situation may be re-described, or that more detail may be added to the description, in a way which completely alters our assessment.

Dissociation Theories

Dissociation theories start from the premise that the self-deceiver knows, suspects, or has a true belief about, what he is doing, and / or the thing that he is deceived about.

They aim to explain how he can nonetheless be self-deceived. To do so, they claim that his knowledge, suspicion or true belief is, in some way, disconnected from his actions, emotions, or other beliefs, which, therefore, it cannot influence. It is, though, connected to wishes or desires, for the self-deceiver (according to dissociation theories) knows (or believes, or suspects) but does not wish to know (or believe, or suspect). The knowledge, suspicion or true belief is disconnected and something else is put in its place. This is the deception which the self-deceiver practices upon himself.

For brevity I will talk about "true beliefs" rather than the full (and longwinded) "knowledge, suspicion, or true beliefs". This will limit my remarks to one variety of dissociation theory; but the remarks can be extended with ease to the varieties of dissociation theories which talk about knowledge or suspicion rather than about true belief.

Dissociation theories disagree about what is disconnected from what. Some of the proposals offered are:

- knowledge that p is disconnected from belief that p (where 'p' is our identifier of what is known, believed, etc)

- belief that p is disconnected from thinking that p

- belief that p is disconnected from (some) actions or emotions

- thinking that p is disconnected from (some) actions or emotions.

These suggestions may all be complementary. They may describe different varieties of self-deception with disconnection as their common theme, and differing with respect to the point at which the disconnection occurs. However, it is perhaps suggestive that the later proposals in my list seem to be developed in response to criticisms which aimed to show that the proposals earlier in the list were untenable.

An objection to Dissociation theories is that they do not explain (or do not explain satisfactorily) how someone makes such a disconnection. For example, one suggestion (Bach [1981]) is that by focussing attention and other strategies for

directing awareness, one may "think" what one does not believe. The "telescope" analogy seems especially apt here. Focussing the telescope on something nearby makes distant things out of focus, and vice versa. One may direct the "telescope" of attention towards one thing, so that all others are outside or peripheral to one's field of vision, "out of focus".

Focussing attention is something we all do: it seems eminently familiar. The metaphor is so apt - too apt, perhaps? Why did that particular metaphor spring so readily to mind? Is it perhaps because the telescope metaphor is one which we already presupposed, the model we habitually use when we think about how we think? So is it not possible, even likely, that what actually happens is that, because we use this particular model to understand thinking, we actually make it come true, shaping our thinking to live up to (or down to) the expectations which the metaphor itself created. If so, then explaining self-deception by this illuminating metaphor is simply finding the evidence we planted in the first place. The metaphor of the telescope seems to fit so well because we presuppose it every time we think about ourselves: we use it in order to understand (perhaps to misunderstand) ourselves. This does not establish that it is true. Every time it does not fit we will discover a strange "failure" of consciousness - which is better described as a failure of the metaphor to fit consciousness.

Suppose we play along with the metaphor for a little longer. Focussing attention seems less familiar and easy to understand in the case of self-deception than it is in other cases. For in the case of self-deception, one is focussing upon something with the aim of not focussing upon something else which is, nonetheless, having to do the work of guiding your strategy of focussed attention. Ignoring something is quite unlike being ignorant of it. You need to be especially aware of something in order to ignore it. It seems as though the thing one ignores must still be "playing upon one's mind". If it is so important, then it may prevent you attending properly to the things you are focussing upon. Still, it might be possible to deceive oneself this way. If something is "playing upon one's mind", in the circumstances described it would do so in an "unfocussed" way.

Self-deceivers, under this description, will have a distracted air, indicating that there is something on their minds. The distracted air may be one of the things which alerts us to the delicate balancing act by which they achieve self-deception.

I have tried to defend Dissociation theories as persuasively as I can. Now I want to say what I think is wrong with them.

1. Firstly, Dissociation theories seem to describe some varieties of self-deception but not others. For instance, they do not describe the brash self-deception of a bigot who is

ready to take on the world with unshakeable conviction. There is nothing fragile or delicate about the bigot's performance, yet I would say that it qualifies for the title of self-deception. For the bigot clings to his belief regardless of whether it is true or false, right or wrong (although, of course, he believes it is true and right, not false and wrong).

2. The description seems incomplete. It does not tell us what the self-deceiver puts in place of the disconnected belief, or how he does the disconnection. We need an account of this too - which I suggest will be in terms of what role the self-deceiver adopts (Role theories) or an attitude or theory which he adopts (Radical Interpretation theories).

3. Dissociation theories explain self-deception by assimilation to something familiar, such as focussing attention. But the penalty for doing so is that instead of self-deception becoming less puzzling, the familiar activity may become mysterious. How do we do it - and do we do it? The explanation could not fail to fit, because it is the filter through which the data is strained before we become aware of it. We "focus attention" - but how do we do that? "Focussing attention", in order to ignore something, seems to me a form of pretence - "pretending that it isn't there" - so that the plausibility of dissociation theories arises from their tacit allusion to a different kind of theory: Role theories, which I discuss in a later chapter.

4. The theories do not show that the belief is sufficiently disconnected to avoid the paradoxes of self-deception. For example, the belief is connected to a wish or desire (otherwise, what motivates the self-deception!), and the wish or desire is connected to the "thinking" - the process of focussing or directing attention onto chosen objects (it must be, in order to guide the self-deceiver's strategy). The belief is supposed to be disconnected from the thinking (or: disconnected from what we think - I doubt that the distinction makes much difference in this context); yet what I have just described is not so much a disconnection as a puzzling connection between the belief and the "thinking", mediated by wishes or desires.

In this case, it may be more parsimonious - and adequate - to give up the claim that there are cases of self-deception, and instead describe them as cases of wishful thinking - as Elster suggests (Elster [1979], p27).

However, we might prefer to classify wishful thinking as a special variety of self-deception. What happens in wishful thinking? A desire suppresses (or represses) the unwelcome true belief and becomes manifest in the desired false or inauthentic thinking, action, emotion, etc.

This description of wishful thinking needs more elaboration: is wishful thinking something one does consciously or unconsciously? In other words, does the "wishful thinker" know what he is up to, or not?

Suppose that someone, S, consciously satisfies a wish by (to borrow Freud's terminology) "hallucinatory wish fulfilment". Suppose that he knows the resulting belief is false: then the self-deception paradox appears again within the description of wishful thinking.

Suppose then that S unconsciously satisfies the wish by "hallucinatory wish fulfilment". In this case there is a false belief but no self-deception, unless we can find some way to hold S responsible for the wish-fulfilment. For unless S is responsible, then he is deceived but not self-deceived. Yet in that case, the alleged example was not a candidate for the description 'self-deceived' in the first place, i.e. Elster would be knocking over a straw man. If we do hold S responsible, then:

- either we have a case of self-deception by negligence (see the chapter on Negligence theories) because he could have done something to avoid having a false belief but failed to do it

- or we have a case of self-deception by radical interpretation (see the chapter on Radical Interpretation theories) because S does something in order to be deceived: he constructs a belief.

Wishful thinking can be as paradoxical as self-deception. If we accept Elster's proposal and make wishful thinking distinct from self-deception, by definition, then we will have "paradoxes of wishful thinking", in place of (or in addition to) the paradoxes of self-deception. These paradoxes of wishful thinking will need to be solved - probably by the same means as the paradoxes of self-deception. For when we explain alleged cases of self-deception as "wishful thinking", the explanation inherits a paradox which formerly arose in descriptions of self-deception.

Elster can be defended by pointing out that the paradox does not arise in a description of wishful thinking unless we introduce the (surely gratuitous) claim that the person who indulges in wishful thinking must have a (masked, veiled, or disconnected) true belief which is inconsistent with the wishful one. But, as I shall argue when I come to discuss Not-Know theories, the claim is also gratuitous when we describe cases of self-deception.

Disconnection theories capture the intuition that the self-deceiver thinks that p because he believes that not-p. This suggests to me that there might be a sort of controlling hierarchy within a person, with a high-level wish and high-level true belief guiding a strategy of self-deception. The idea here would be that self-deception can work because one's consciousness of oneself and others always exists at a low level in the hierarchy: that one never knows what the higher levels in the hierarchy are doing (though one may be able to make educated guesses from their manifestations in behaviour). Conscious beliefs, desires, intentions, and so on, would always be the lowest and least significant level: they would be manufactured by the activity of the unconscious ones. But this, of course, is pure speculation, not grounded in any kind of empirical testing. I leave it to psychologists to sort this sort of question out.

The gist of Dissociation theories is that they tell us the self-deceiver believes (knows, suspects) that p but acts as if, thinks as if, feels as if, he believes that not-p.

They say that this can be explained as a disconnection. But, I have argued, it seems more like a strange connection than a disconnection. How is it achieved? The activities mentioned (focussing attention, etc) seem more like tactics than strategies. So what is the unifying strategy which integrates

all the tactics into the overall project of self-deception? The "disconnection" seems to play only a very minor explanatory role, and the major part of explaining self-deception will occur not in describing the disconnection, but the things the self-deceiver does in order to put something else in place of the "disconnected" belief. One wonders if the disconnection may not be a symptom of self-deception rather than a cause - if it occurs at all; and who is to say that it does?

The dissociation occurs when the self-deceiver fails to use the belief he has. He does not use it to guide action, or to interpret data, and he "steels" emotions in order not to feel the emotions which the belief evokes.

However, if we describe self-deception in this way, there is a problem. The problem is that either the belief is redundant (it performs no detectable function, so there seems little reason to propose that it exists), or the self-deceiver's actions, feelings, and interpretations are explained by proposing that the belief motivates or in some way influences them. If the latter is the case, then the belief is not disconnected but connected, in a strange way, to actions, feelings and interpretations. In that case it seems that the self-deceiver is not really deceived, since he has the belief.

We may prefer to describe self-deception by using an option different from Dissociation theories. We could argue that the self-deceiver does not have the "unwelcome belief". He may have an unwelcome item of information but he does not believe it. The self-deceiver uses actions, interpretations and emotions in order not to have the unwelcome belief.

At this point I might as well cement together two ways of talking about self-deception, both of which crop up independently in discussions of self-deception. Sometimes self-deception is described as avoiding an unwelcome belief (which the description usually also characterises as true), and sometimes it is described as aiming at a favoured belief (which the description usually characterises as false). These two descriptions are compatible and seem to me to belong together: the self-deceiver aims for a favoured belief in order to avoid an unwelcome belief; he avoids an unwelcome belief because he aims to have the favoured belief.

Sometimes we use theories which we do not believe as "instruments" to achieve some aim. Usually these instruments do not interfere with our beliefs; usually the use of the instrumental theories is guided by the beliefs. However, an instrumental theory might contradict the belief which guides it; and using the theory might "mask" the belief by preventing

the user from being conscious of the belief. So this seems like a way in which one might be self-deceived.

This scenario can be illustrated with an example. Take some of the examples which are common in the literature about self-deception, and let us test them against this proposal. Bear in mind my claim, made in the chapter about examples, that descriptions of self-deception incorporate theories about self-deception. One must therefore treat the examples with caution.

1. Suppose that Mr Dread has evidence that his son is a criminal. However he invents a different interpretation of the evidence and thereby avoids the painful consciousness of the son's criminal activity. But this seems to be an example of someone avoiding a belief rather than being deceived despite having the belief. He may originally have interpreted the evidence in such a way as to gain the information that his son is a criminal. But he does not believe it. We all have plenty of items of information which we acquired by forming interpretations, but which we do not believe. We may well know that, for example, the British economy is coming out of recession, according to the Chancellor of the Exchequer; but we may not believe it. Having done the interpretation does not guarantee that we believe the result. Mr Dread did the interpretation but did not believe the result; not liking the result, he found some other interpretation instead. So, I

suggest, this is not an example of someone being self-deceived despite having a true belief.

2. Another example is an instance of "protesting too much". Mrs Dread thinks (and says) that her son is not a criminal in order to avoid thinking what she really believes, i.e. that he is a criminal. She apparently needs to argue the case constantly, even when no-one else raises the issue. It seems that her arguments are not defences against what other people say or think, but against her own belief, against which she is desperately fighting.

However, surely denial is not, in most cases, evidence of belief: if Mrs Dread denies that her son is a criminal, that is surely evidence that she does not believe it. "Protesting too much" is a sign of being aware of an interpretation which one is refusing to believe (or which one does not want others to believe).

Once again, I suggest, we do not need to propose that the self-deceiver believes that her son is a criminal (assuming that Mrs Dread is self-deceived). She can be re-described as preventing or deferring the belief rather than having an unwelcome belief and "masking" it. This is the "shaving paradox" again: shaving is a way of preventing unwanted surplus hair; one does

not need to preserve a beard, for example, in order to motivate the act of shaving.

3. The self-deceiver may have a belief which circumstances later make unwelcome. For example, a doctor may believe that an x-ray photograph which shows a shadow on her patient's lung is evidence that the patient has cancer. Later on she discovers that the x-ray photograph is of her own lung. But the conclusion, "I have cancer" is unwelcome; so she starts to deny that the x-ray photograph provides evidence of cancer. In addition she may use her medical knowledge to guide her self-deception, pre-empting other evidence that she has cancer ("the only reason I have a low white blood cell count is because I had a few stiff whiskies before having the blood test," she argues, while arranging to be too busy to have the blood test again).

The doctor not only has the knowledge which undermines the self-deception, she even uses it to protect the self-deception from refutation. However, one may use an interpretation which one does not believe, in order to pre-empt arguments with which one does not agree. I am told that during the Second World War the British Ministry of Defence used astrologers to predict what Hitler was planning to do next - not because people at the Ministry of Defence believed in astrology, but because Hitler did. In the same way, we can argue, the doctor uses the

medical knowledge not because she believes it, but as an instrument, because it enables her to predict and disarm arguments against her own belief that she is perfectly healthy. Yet, she used to believe the arguments which she now denies: does she not still believe them, deep down, even though she is not admitting it to herself?

If this example is plausible then perhaps we have found an instance of self-deception which incorporates the "unwelcome belief" it sets out to deny. Notice, though, that in this case the medical knowledge is actually a complicating factor in our description of the self-deception: someone who had less medical knowledge might have found it much easier to be self-deceived about their state of health - "doctors? What do they know about it?" - for the doctor has access to a more detailed and systematic body of (unwelcome) knowledge. So, I argue, the "unwelcome belief" is not an essential part of self-deception, but a complicating factor. It can be accommodated within the account of self-deception which I am going to propose later on.

4. My last example is the case of the bureaucrat's in-tray. Suppose a high-ranking civil servant sees a document in his in-tray, labelled in bold letters: 'Evidence Of Corruption In The Civil Service'. This is unwelcome news to him, and he makes sure that he does not read it. But if he does not read it he cannot find out whether or not the document concludes, "we

could find no evidence of corruption". Let us suppose, he avoids reading the document because he knows what it will say and cannot refute it: "deep down" he knows already that the document is not going to vindicate the civil service, but he does not want to think about it; if he did, he might have to do something about it. The very thought of having to do something about it arouses most unpleasant emotions, so he prefers to leave the subject of corruption hypothetical and not enquire into it too closely. He avoids disturbing the interpretation he currently uses ("everything is above board") and, however strongly he suspects, or even knows the truth, he avoids "the glaring truth".

I provisionally accept this example which supports the claims of Dissociation theories, in order to see where it leads us. The question is, what does the bureaucrat do instead of using the unwelcome belief (or suspicion, or knowledge) directly? How is the disconnection done? He uses his belief (or suspicion, or knowledge) in order to guide another interpretation, a pretence. To pursue this element of self-deception we need to supplement the Dissociation theory with a theory about pretence (a Role theory) and a theory about interpretation (a Radical Interpretation theory).

Dissociation theories are an advance upon Schism theories. They do at least offer some explanation of the connection

between the two sides of the "schism". I suggested that there must be such a connection. I suggested that the connection could be explained as the mediation of a desire: an unwelcome belief, plus the desire which makes it unwelcome, leads the self-deceiver to use a different interpretation, a pretence. This is not a disconnection but a "strange connection".

Role theories offer an advance beyond Dissociation theories by emphasising this element of pretence which, I suggested, is needed in order to give Dissociation theories their plausibility. Role theories also allow us to spell out in a bit more detail the theme that the self-deceiver is "too well integrated". The self-deceiver is too well integrated because he never steps out of the role he has adopted, and so he can never engage in the interpretations which might make that role untenable. The role "masks" those other interpretations and the conclusions to which they lead.

Role theories promise to explain what the self-deceiver puts in place of the "disconnected" belief. They also mesh precisely with the description which Dissociation theories virtually force upon us: the self-deceiver acts <u>as if</u>, thinks <u>as if</u>, feels <u>as if</u>, such-and-such is true. Simulation and dissimulation ("as if") are at the heart of Role theories. The next several chapters set the scene for a discussion of Role theories.

Data Is Not Evidence

There are two distinct sorts of thing both of which can be referred to by the word 'evidence'. This could be confusing since I want to talk about both of them; so to avoid confusion I shall reserve the word 'evidence' for one of them and use the word 'data' for the other.

I distinguish data from evidence, as follows. Evidence is "that which makes evident"; data is the raw material from which we can get evidence by a process of interpretation.

Below I give two examples to illustrate my use of the words 'data' and 'evidence'. The first example comes from the data processing industry, from which I derive my use of the word 'data'. The second example is taken from the law courts, where the use of evidence is so important.

First Example

A gas company stores its computerised account records on magnetic tapes. However nobody at the gas company can read the tapes just by looking at them. In order to retrieve the stored

records, the tape must be loaded into a computer, and a computer program must be used to translate the records into a form that someone can read. The patterns of magnetism on the tape are, in the company's view, uninterpreted data. The program is used to convert the data into information. But until someone reads the information, it does not inform anybody.

So the program is used, lets say to produce a gas bill for sixty pounds. The gas bill is sent out to a customer. Seven days later the customer's reply is received. He disputes the bill, claiming it is too high.

The people at the gas company are perplexed. How can the customer dispute the amount owing, when he has the evidence, namely, the gas bill. They suppose the reply must be due to wishful thinking on the part of the customer.

"Evidence" (that which makes evident) is relative to its intended audience - it is that which makes something evident to somebody. The gas bill was written in English (the sort of English used by gas companies). The customer did not understand it, mainly because he speaks Portuguese and not English. For him, the bill was uninterpreted data. He could interpret it sufficiently to recognise that it was a bill, however, and he found an interpreter to translate it. The

translation did not make it evident to him that he owed the gas company sixty pounds. What it makes evident to him is that the gas company has made a mistake - or something worse than a mistake, an attempted deception. For he has been reading his gas meter, and calculates that he only owes twenty pounds.

Had he used the same interpretation as the gas company, then he would have arrived at their conclusion, namely that he owes sixty pounds. It would have been evident, but it would have been false.

The people at the gas company have forgotten that their evidence is fabricated. Normally to say that evidence is fabricated means that it is counterfeit, "planted" evidence. That is not what I mean. It refers to what I would call "planted" data. In my sense of 'evidence', all evidence is fabricated, i.e. constructed. The gas company's evidence is constructed from stored data by programmed computers. They have not questioned the construction process. There might be a fault in the computer program, for instance. There may be more than one program, more than one way of interpreting the data. They may be using the wrong interpretation. Let's suppose they are. If they had retrieved their information using a different program, they would have noticed that the bill was only estimated: the customer's meter was never read. In this case

the data was "fabricated" or "planted": lacking a meter reading, they substituted an estimate in its place.

There are several lessons to be drawn from this example:

a. evidence is constructed by a process of interpretation
b. evidence is relative to an audience, and in particular to the process that audience uses to construct the evidence
c. evidence is constructed from data
d. uninterpreted data does not make anything evident
e. what counts as data is relative to a process of interpretation. So, for example, the gas bill was the result of the gas company's interpretation of their data: for the company, the gas bill is evidence. But to the customer it is uninterpreted data until he interprets it (or, in this case, finds an interpreter to translate it for him: he must then interpret what the interpreter says, in order to arrive at some evidence). Evidence from one process may be data for another process.

Given points a through e, it is easy to understand how a self-deceiver might fail to be convinced by "the evidence". For he may not do the interpreting that others do. If no processing takes place, then no evidence is produced. If different processing takes place, then the resulting evidence may be different too, making different things evident (e.g. that the

gas company has made a mistake, instead of "I owe sixty pounds").

Since evidence (in my sense of evidence) is constructed, we may choose not to construct it, or to construct different evidence out of the same data.

Second Example

A criminal case is being heard before a court of law. The court usher places exhibit A on a table before the jury. The jurors look at exhibit A, which appears to them to be an ordinary house brick. The exhibit is "brought in evidence", but as yet they do not know what it is supposed to make evident. For them, its role in the case is unfathomable: it is uninterpreted data.

The data becomes connected with evidence by virtue of its role in the process of interpretation. The prosecution brings witnesses to give evidence. They tell a story about the brick: they claim that the defendant used the brick to break a jeweller's window, so as to steal jewelry.

Is there now evident that the defendant is guilty? I suggested that the witnesses gave evidence: perhaps it is better to say that they testified. For next the witnesses for the defence

come along and tell a different story. According to this story, the witnesses for the prosecution were seen to break the window and then hand the brick to the defendant, before arresting him as a thief. The defence lawyer points out inconsistencies in the testimony of the prosecution witnesses. The defence lawyer is using the testimony of the prosecution witnesses as data, not as evidence. He interprets it as false evidence, and what it makes evident is not what the prosecution claims. The same data, interpreted in different ways, leads to different conclusions.

I use the word 'interpretation' to refer to (a) a process, and (b) the products of that process. Evidence is such a product. The process is one by which data is used to construct evidence and by which "conclusions" are "drawn from" evidence. I put these expressions in scare quotes because I think that it is misleading to call these items "conclusions" as though they always came at the end of the process, and it is a mistake to suppose that the items are always or even often derived from the evidence.

Evidence is constructed, not found. Data is found, not constructed. There is always an element of invention in the construction of evidence: the invention may be justified - for example, when we tell a story about something, and the story turns out to be true. There is also an element of risk: the

story may turn out to be false, so that we become aware that the story was an invention, a fabrication. We may say that the evidence was misleading; data cannot mislead, for it cannot "lead" either. We construct our true stories in just the same way as our false stories, so clearly they are inventions also: inventions which happen to be true.

By testing the story we may be able to show that it is false. We cannot show that it is false without putting it to the test.

The process by which evidence is constructed may be altered by our choices. The process cannot be performed if resources are not available for it: time, effort, data, for example. If we do not find time to think about the data, if we do not make the effort to do so, or if we do not find the data to begin with, the process cannot take place. If the process does not take place, the evidence is not produced. Without the evidence, nothing is made evident. If nothing is made evident, there is nothing available to believe. The evidence cannot "force us to conclusions" if it does not exist.

If the process is performed, then it may be starved of resources: we may not be able to think long and hard enough and well enough, if we are tired, ill, harassed, emotionally distraught, and so on. And sometimes it may suit us to become emotionally distraught about some topics - those where we do

not wish to think through to the conclusions. We know in advance how to guide the process since, far from concluding the process, or being "concluded" from evidence, the "conclusions" are often the start of it - as I shall argue below.

If the process is distressing - producing ideas that we find hateful, for instance - then it may take great efforts to carry on with it. We may "face up to it" or "give in to it" - both expressions indicating that such processes do not take place wholly involuntarily, even if they are not wholly voluntary either.

We may also perform the process negligently: failing to apply epistemic norms: thinking "sloppily"; and so on.

The process may be altered by our aims and wishes: as I suggested above, the process is an exercise in story-telling, which is at least partly voluntary.

Data cannot force us to conclusions, for uninterpreted data is not evidence: it does not make anything evident.

Interpretations of data - evidence - cannot coerce us into believing anything, because all evidence is a product of one or another process of manufacture: and since this process is likely to take time, effort, patience, ingenuity,

resourcefulness, commitment, endurance, and so on, it is something we can make choices about: to do it or not. Since the process may also involve us in great anguish or great pleasure, it may take lots of willpower to pursue it, or to desist. We may need to "face up to it" or try not to "give in to it"; this kind of venture is quite unlike the mechanical sequence of events which is called to mind when we are told that evidence is "compelling", that it "forces us to conclusions", "inclines the judgement" to one side or the other of an argument, and so on. I have more to say about these metaphors, later on.

Does anything compel us to choose one such process rather than another? We may fall victim to force of habit; we may be more or less addicted to one particular process; but in principle and often in practice we can choose. Sartre considered "bad faith" to be a refusal or denial of a vertiginous freedom which brooks no denial and allows no refusal; and bad faith has a lot in common with self-deception, if it is not identical to it. We draw back from the ways of interpreting of which we disapprove because it is too easy to become enmeshed in them. Thinkers in the Western liberal tradition do not as a rule wish to be open-minded about Nazism, for example: do not wish to understand what it would be like to hold those views, because to be open-minded is to be open to persuasion; understanding is only a step away from sympathetic understanding, which is

only a step away from collusion, collaboration, fellow-travelling, and eventually participation. We do not wish to picture what it would be like for us to share the views we despise, for that comes uncomfortably close to seeing if the cap fits, becoming what we pictured ourselves being.

"See it from my point of view" says the recruiter; but to do so is sometimes to be lost. To see it from that point of view is to think with those patterns, to make those things obvious, to become immersed in it, with the possibility that we may never detach ourselves from the process for long enough to make a decision to escape it or not.

Once we take notice of the construction of evidence, we can see similarities between self-deception (by which one makes something which is false seem obvious to oneself), and some kinds of (not self-deceptive) innovative enquiry. If evidence were coercive in a way which prevented self-deception, then it would also prevent these (very valuable) kinds of enquiry.

Radical innovators in enquiry can face problems similar to those of a radical self-deceiver. It is worth noting some of the similarities.

By "radical innovation in enquiry" I mean such things as the production of hypotheses which are wildly at odds with those

currently accepted, the well-established hypotheses which are widely used and familiar, and which are taken to be commonsensical and perhaps even indisputable. The radical innovator disputes them nonetheless.

Ways of thinking become habitual as they become established; habitual ways are also likely to be easy ways, which makes their results (the "conclusions arrived at") seem obvious since so little effort is needed to arrive at them.

Breaking established habits of thought can be difficult, needing plenty of effort, endurance and imagination; when we do break them, the results of a new way of thinking are likely to seem implausible because they are counter-intuitive. This should be no surprise since the intuitions are formed by the established ways of thinking.

There will be very little evidence available to "support" the new hypothesis, because the available evidence was produced by interpretations of data generated by the established ways of thinking. These interpretations of data may have been created by generations of people all working within the established ways of thinking. No similar amount of work has been done using the new approach; the amount of work which might have to be cast away or done all over again may be daunting.

The established ways of thinking will also be closely woven into established ways of doing things: a wholesale change is impossible and a piecemeal approach will lead to constant conflicts with "unreformed" ways of doing things; the inevitability of such conflicts will mean that inconsistencies will be sustained for a very long time.

For the innovative enquirer, it will be easy to relapse into the established ways of thinking, "losing the thread" of the unfamiliar and difficult new approach.

Innovative thinking is therefore of necessity sustained by the hope of things to come and not by the available evidence, hunches, intuitions, and so on.

Despite these difficulties, innovative thinking is possible. If the innovative enquirer can resist the reign of the obvious by effort, endurance and imagination, then so can the self-deceiver; and sometimes it will be difficult to tell them apart, the towering genius on one hand and the self-deceiving crank on the other. There may even be something of both in a single person. One might say, though only to emphasise a point, that lucky cranks turn out to have been pioneers, advancing the goals of enquiry; Unlucky cranks turn out to have been self-deceivers.

Self-deception is possible because innovative enquiry is possible: both can take place only because we are able to resist the tyranny of the obvious and defy all evidence, intuition and authority. We may find it mystifying that self-deceivers can be oblivious to "the" evidence, immune to what seems obvious to us (and surely must also seem obvious to them, we may think): but the innovative enquirers also do just that. It is possible because evidence cannot compel us to accept conclusions.

In this chapter I have argued that data cannot compel belief because it does not make anything evident, and evidence cannot compel belief because it is a product of something which we do and which we can therefore do differently. Beneath the "magic button" of evidence are the processes of interpretation.

In the next chapter I discuss a process which we are often cautioned against performing: leaping to conclusions. I shall argue that although we are right to be cautious ("look before you leap"), the "leap" is an indispensable part of enquiry without which we cannot gain evidence ("leap in order to look").

Leaping To Conclusions

"Leaping to conclusions" is inventing a story and then believing it - a useful skill for a self-deceiver to have! But we only leap to conclusions which seem to us obvious. Being obvious or evident is not something which items such as sentences or beliefs have by virtue of themselves alone: it is relative (being obvious to some specific person in some specific circumstances) and it is conferred by a process which makes them obvious. What is obviously true may not be genuinely true.

Data and evidence are distinct from each other. Data are the raw materials which are used to make a story. 'Evidence' names the role of data which are incorporated into such a story. 'Conclusions' names those other items in the story, which are invented to explain the evidence.

When no data is available, there is nothing to tell a story about. When all the data is available, there is no room for interpretation and so no occasion for some data to play the role of "evidence" from which we may draw "conclusions" about missing data. So we need is some available data, and some data

missing, to have both evidence and conclusions. But then the evidence cannot compel us to believe the conclusions, because it is only by virtue of our inventing the conclusions that data can play the role of evidence for them.

It is worth emphasising the importance of invention or fabrication in the process, because otherwise it may seem that evidence is a natural product and that we only need to find enough of it in order for it to "compel" us to arrive at "obvious" conclusions. Self-deception would then have to be avoidance of evidence. For if the self-deceiver encountered "compelling" evidence the deception would be destroyed. Self-deception which is sustained when the self-deceiver encounters "compelling evidence" would be as mysterious as an encounter between the irresistable force and the immovable object.

Negligence theories describe self-deception as a deviation from epistemic norms - norms for seeking truth and avoiding falsehood. But it may be more accurate to say that these norms are a deviation from self-deception: the norms have a historical development; enormous efforts have been made to create and improve standards of enquiry, a methodology for gaining truth. Self-deception, by contrast, has no methodology, no standards, and seems relatively effortless. So self-deception may belong to the normal pattern of behaviour. Instead of counting self-deception as a distortion of a truth-

regarding process (enquiry), we should count enquiry as a development of a process which is not primarily aimed at truth.

Seeking truth is one way of acquiring beliefs, but it is not the only way. When we regard it as the only way, we are bound to run into paradoxes with regard to self-deception. For to fit self-deception into the mould of truth-seeking, we have to explain either how someone can aim to gain false beliefs by seeking truths, or else how someone can gain beliefs by seeking falsehoods (for to believe something is to believe that something is true: so seeking falsehoods will lead to disbelief, not belief).

We are often cautioned against such things as jumping to conclusions on insufficient evidence, against partiality and against wishful thinking. There is no methodology for jumping to conclusions (to take one example); no instruction manual tells us how it is done. Yet there would be no need for a rule forbidding it unless someone were tempted to do it. Since there is no instruction in the art of jumping to conclusions, it must be something we do naturally.

With this ability, all a self-deceiver need do is to jump to the desired conclusion, taking care ever afterwards not to assess that conclusion with regard to the norms of enquiry. However, just because jumping to conclusions is something we do

"naturally", perhaps it is not so easily controlled: perhaps it is difficult to make sure that one jumps to the conclusion one wants to believe, and not to some other. For people usually jump to "the obvious conclusion"; whereas often the self-deceiver believes what seems (to other people) not at all obvious. This is one of the puzzling things about self-deception: how someone can be deceived "despite all the evidence", when the truth is "so obvious".

'Obvious' is relative: what is obvious to me may not be obvious to you; what is obvious to one person may differ according to time and place; what is obvious at the end of a process of thinking or a course of action may not have been obvious at the start. Being obvious is not a property which the sentences or beliefs have by virtue of themselves alone: it is the outcome of a process which makes them obvious.

There always is such a process whenever something is obvious; things are obvious because something makes them obvious, and not otherwise.

Sometimes the process may be so trivial that we do not notice it. For example, the sentence, 'a spade is a spade' seems to me obviously true. The processes which make it obvious to me include recognising that the sentence is a sentence of English; scanning the sentence with my eyes; interpreting what I see;

deciding that the first occurrence of 'spade' refers to the same thing as the second occurrence of 'spade'; and so on. But I do not usually pay attention to these processes. Usually I do not notice them at all.

The self-deceiver would need to make the desired conclusion obvious to himself or herself, indeed "blindingly obvious", since it is intended to blind the self-deceiver to other possible conclusions - otherwise he would not know to which conclusion he means to "jump".

There are ways to make false things seem obviously true; e.g., by the interpolation of irrelevant information ("the angle on the right is called a right angle, what's the angle on the left called, a ...?" in this case the location of the angle is emphasised as though it has some bearing upon the name of the angle); or by the application of an irrelevant rule ("if a man from Poland is called a Pole, what's a man from Holland called, a ... ?" In this case a rule about rhyming is introduced to link 'Pole' with 'Hole'). Though these are trivial examples, they can be compelling in some circumstances e.g. in quiz games when people are under pressure to answer quickly - and have to "jump" to the conclusion. The primary way of making something seem obvious is to tell a story which makes it obvious.

The pressures of a quiz game are mild compared to the pressures of everyday life: so that in practice we are often obliged to jump to conclusions, against the norms of enquiry. Indeed, there is always a "jump" beyond what can be derived from the data available to us, unless the case is a purely logical derivation. Not only is the "jump to conclusions" natural, it is also the only way we can arrive at conclusions at all. The rule against jumping to conclusions is misdirected: we have to jump to conclusions - it is the only way to gain conclusions; but we should be cautious of relying upon conclusions we have jumped to without further testing.

How do we jump to conclusions? Take the legend of Gelert as an example. In this story Prince Llewelyn goes out hunting, leaving his hound Gelert to guard his child. On his return the Prince finds Gelert covered in blood yet apparently unharmed; the child is missing. He jumps to the conclusion that Gelert has killed the child. i.e. he invents a story to explain the data. This story is sheer fabrication. It happens to be false, but it would still be sheer fabrication even if it were true.

Llewelyn invents a story which incorporates a few items of data. By virtue of being incorporated in the story in the way they are, the data become evidence for the "conclusion" that Gelert killed the child, though they fall far short of

"compelling" evidence. Calling it a "conclusion" is grossly misleading, since far from being a conclusion it is the invention from which everything else commenced: Llewelyn had to invent the "conclusion" before the data could be given the role of "evidence for the conclusion".

I shall have more to say about "evidence" and "conclusions" shortly.

Having invented the story, Llewelyn does not assess it with regard to considerations of truth. He does not try to invent any rival stories which would incorporate the data differently. He gives himself no alternative to this single story. He does not seek any more data to test the story; he does not seek the body of the child, or look at the scene where his story takes place. The story arouses powerful emotions in him, and he allows them to draw him into immediate action. Acting upon his interpretation of the data, he kills Gelert with his sword.

The events happen more easily because Llewelyn seems not realise that the situation as he sees it is his own invention. He has invented the story, but he does not seem to notice that it is what he tells himself to explain the data: it seems to him to be something which "appearances" themselves tell him. He does not distinguish the means by which he sees the situation, from the situation which he sees; what he sees is

that Gelert has killed the child - for this is what the story tells him. The appearances mislead him because he has interpreted appearances in such a way as to mislead himself.

Yet he is not a self-deceiver. His intention is to gain truth, but he "falls away from the norms for seeking truth". He acts upon the first hypothesis he comes across, without testing it further and without seeking alternatives. The conclusion does not appeal to him: he does not jump to a conclusion that suits him, as a self-deceiver would (I am excluding those peculiar explanations which might show that e.g. at some deep level of his psyche he desired to punish himself and so the conclusion really did suit him, etc). If Llewelyn had been a self-deceiver he would believe what it suited him to believe. He would not believe that his child was dead, unless the evidence was unavoidable; he would not believe that Gelert had killed the child.

Suppose that the story did not proceed in the tidy way of the legend: Llewelyn does not find the body of a huge wolf; and the child is dead. It does not suit Llewelyn to believe that Gelert killed the child. So he fabricates a story which suits him and which is compatible with the available data: he argues that a huge wolf came and killed the child; and that Gelert fought the wolf, which slunk away to die somewhere else. This

conclusion suits him, so he does not seek to test it or to create rival stories.

This scenario fits three descriptions of self-deception given by "Not-Know" theories: it is negligent (Negligence theories), since it does not assess the story with regard to truth considerations; it is radical (Radical Interpretation Theories), since it invents a story not supported by the circumstances; and Llewelyn simulates the role of someone who genuinely believes (Role Simulation theories).

In discussing the story of Gelert I have tried to circumvent something which makes self-deception more puzzling than it need be, namely the supposition that evidence is coercive: "compelling evidence" "forces" us to believe the "conclusions" it "supports". Here almost every word is a misnomer.

The conclusions do not conclude the process; they are the beginning of the process, not the end. The evidence does not lead to the conclusions, i.e. the process does not have evidence as an input and conclusions as an output: the evidence and the conclusions are both output (and therefore the evidence cannot "lead to" the conclusions, nor be "foundations" for them). The input is data. 'Evidence' is the name of a role which data plays within a story. 'Conclusion' names any element of the story which is not an item of data; calling it

a 'conclusion' picks it out as something which the data "supports" by virtue of the data's role as "evidence" within that story.

The effect of the sentence is to make self-deception puzzling because self-deception is made to seem like the irresistable force (evidence) meeting the immovable object (the self-deceiver's belief).

Putting things in this way forces us to say things like "the self-deceiver distorts the evidence" rather than "the self-deceiver uses the data to construct a different sort of evidence"; we are obliged to say, "the self-deceiver is selective about evidence", as though there were something wrong in being selective; everyone is selective about evidence because some data has no role to play in the things we want to make evident, being irrelevant.

Suppose that instead of having to interpret a few data, Llewelyn could actually watch the scene unfolding, e.g. he watches through a telescope, being too far away to intervene. He sees Gelert attack the child and kill it. There is no sign of a wolf. Now he has so much data that there is very little room for creative interpretation: the story is so detailed that there is very little to add, and so there is not much opportunity for self-deception.

To deceive himself now would require great efforts: he would need to persuade himself that this original interpretation of data (watching the scene through the telescope) is misleading: perhaps he is hallucinating; perhaps there is something strange about the telescope; could it be that what appeared to be Gelert was actually a wolf? and so on.

So "evidence" is most "compelling" when there is little opportunity for interpretation: when there are no missing data and so no need to "draw conclusions": we already have the whole story.

What is obviously true may not be genuinely true. Since the obviously true conclusion may well be the outcome of a flawed manner of interpreting data, someone seeking truth may need to set aside the obvious conclusion, resisting the force of habit and the comforts of the familiar ways of doing things in order to arrive at a novel interpretation (which may or may not be true). If a truth-seeker can resist the obvious conclusion, so can a self-deceiver.

Belief At Will

Williams [1973] describes belief in terms of its intimate connection with truth:

> belief aims at truth. (p136)

He argues that, in consequence, it is necessarily - and not just contingently - true that we cannot "believe at will". Later on I argue that Williams' description of belief is not justified, and so his conclusion (that we cannot believe at will) does not follow. But for now I will provisionally grant him the authority to legislate about the grammar of the word, 'believe'. Suppose he is right, and we cannot believe at will.

Williams argues that if I recognise something to be true, I thereby believe it (by the definition of 'believe', for "belief aims at truth").

The word 'recognise' is particularly felicitous in this context. For example, government A may <u>recognise</u> that a particularly bloodthirsty dictatorship, B, has come to power in a neighbouring country. A recognises that B is now the government of the neighbouring country. But A does not accord

B any diplomatic rights: so in the diplomatic sense, A does not recognise that B is a government. Likewise, we may "recognise" that something is true without according it any "diplomatic rights": describing something as true does not prescribe what we are to do about it.

Hare [1952] distinguishes "descriptive" from "prescriptive" uses of moral expressions; e.g., someone may describe an action as "bad" without accepting that therefore he ought not to do it. In this case 'bad' is descriptive, not prescriptive.

We can also apply the descriptive / prescriptive distinction to the use of expressions like 'true' and 'believe'. Someone - lets call them "S" - might recognise that something is true (as a description) - without recognising it (as a prescription of what one ought or ought not to do, e.g. believe it). The "belief" (according to Williams' definition of 'belief') may have no consequences for the way S behaves, thinks, speaks or feels. S has accepted 'true' as a description without according it any prescriptive force. S does not have to think about what he "believes"; nor need he use it to interpret anything or to guide action. He does not even have to feel that it is true, if he can help it, nor incorporate it into any other aspect of life. He does not have to use it to do any of the things for which we might want to use beliefs.

This starts to look like a rather "thin" sense of belief. Beliefs may not be amenable to will; yet they may be irrelevant to all the things which S can do at will, such as:

- hypothesising, guessing, interpreting, connecting ideas, inventing metaphors

- performing actions

- asserting.

If S's emotions are linked to the things which he can do at will, then S's emotions too may be independent of S's beliefs.

In this situation the "belief" is disabled: it is, in effect, consigned to "storage" in S's memory along with all the theories, hypotheses, stories, dreams, fantasies, and so on, which he knows about but does not believe. S can invent any hypothesis he cares to, including ones in which the 'belief' is false, and can act in every respect as though the hypothesis is true.

It seems that S will then be self-deceived, or something very like it. For each variety of theory about self-deception can furnish a description of self-deception which applies to S. For instance:

- the belief is "disconnected" from action, assertion, thinking (e.g. inventing hypotheses, connecting to other beliefs, etc)

  . this terminology is typical of Dissociation theories

- the belief is removed from its role, and pretence or simulation fulfils the role of belief instead

  . this terminology is typical of Role theories

- S neglects epistemic norms, he "falls away" from them

  . this terminology is typical of Negligence theories

- S engages in radical interpretation

  . this terminology is typical of Radical Interpretation theories.

We can also see why non-cognitive theories should attribute relatively little importance to cognitive expressions like 'believe' (when defined with respect to its intimate connection to truth, as Williams defines it). For if beliefs can be disabled, then the important parts of self-deception, its "active ingredients", can all go on independently of belief.

We can even posit a "schism" between a part of S in which the belief resides and another part which performs all those things which he does at will (as in Schism theories of self-deception).

Since every theory about self-deception seems to agree in furnishing a description of self-deception which fits the case of S, I think we may safely say that what I have described is a prototypical case of self-deception, or something very like it.

I have some supplementary remarks to make about the case of S. Notice that S "knows" what he is up to, but that this knowledge is wholly ineffectual, unconsidered, and irrelevant to what he does. S is "living a lie". Notice also that S's rejection of the prescriptive force of truth need not be wholesale: he may pick and choose when to reject it and when to accept it.

We may argue that S is behaving irrationally. That does not mean that he does not have intelligable reasons for doing so.

S's instrumental understanding (or misunderstanding) displaces (or replaces, or devours) S's truth-regarding belief. A misunderstanding can be a most effective instrument for "steering around the evidence", especially if it is designed to do so. The "belief" has no role to play in self-deception, because the self-deception can be guided and controlled by use

of a false belief, a misunderstanding. This removes one reason for claiming that a self-deceiver "must know what he is up to", or "must know what he is deceived about". Truth - or true belief - has no power which compels us to accept it as a prescription of how we ought to behave with regard to it. A mere knowledge of facts does not at all decide for us what we are to do about those facts.

Although all the theories apply to the example, they are not all equally worthwhile. For example, all Avoidance theories (Sschism, Disconnection, and the "avoidance" version of Role theories) include the false claim that a self-deceiver must, of necessity, have a true belief about what he is doing. But it is not a prerequisite of self-deception. If S does have such a belief, it is coincidental to the processes of self-deception. Such a belief may complicate our description of a particular instance of self-deception, since the true belief may not be wholly de-activated: it may interfere in the processes of self-deception. But it is not a necessary feature.

I wrote rather glibly of the "prescriptive force of truth" - what is it? The prescriptive force of truth is that we should believe it:

> whatever else one does with a truth, believing the proposition that expresses it is the first and most

> fitting thing to do with it - before we start deploring it or trying to alter it, for example. The connection between belief and truth is that belief is appropriate to truth; it is proper only when it is of what is true, and only intelligable, therefore, when it is of what could be true. (A. Phillips-Griffiths [1967], p140)

According to Williams' description of 'believe', though, we cannot believe at will, so it cannot be something we ought to do.

I shall now argue against Williams, that belief at will is possible, and that belief can be defined functionally in terms of the things which we can do at will - in its relation to action, emotion, interpretation, perception, other beliefs, wishes, etc. I do not aim to provide a detailed functional definition of belief in this thesis. However I shall have things to say about it indirectly.

If I am right in claiming that it is possible to believe at will, then Williams is wrong. In the next chapter I shall explain what I think is wrong with his arguments. I provisionally gave the word 'belief' to Williams. Now I want to take it back again.

Pretence Revisited

If belief at will is possible, then it seems we cannot distinguish beliefs from merely pretending.

Let us see how well pretence stands up to Williams' criteria for belief.

> Truth and falsehood are a dimension of an assessment of beliefs as opposed to many other psychological states or dispositions. (Williams [1973], p137)

Pretences too can be assessed for truth and falsehood, e.g. the pretence that I am Napoleon Bonaparte is false.

> If a man recognises that what he has been believing is false, he thereby abandons the belief he had. (Williams [1973], p137)

One need not abandon a pretence just because one recognises it to be false. However, recognising that it is false does restrict the pretence. If an actor playing Othello upon the stage takes time out to assert that he recognises that he is not Othello, he is commenting upon the pretence (that he is

Othello) and not pretending. I shall say more about this shortly. Williams writes that,

> to say: 'I believe that p' itself carries, in general, a claim that p is true. (p137)

To say 'I pretend that p' carries no such claim.

So pretence does not seem much like belief. But that is because there are many forms of pretence, and belief is only one of the forms of pretence. Belief is pretence, but it is not "mere" pretence. "Mere" pretences are performed only in restricted circumstances and for limited purposes.

For example, daydreaming is a form of pretence, elaborating an answer to the question, "what if p were true?" But this pretence is not used to guide actions, as beliefs often are. It is restricted.

A liar uses a pretence in order to guide the actions of others, not to guide his own actions. The pretence is restricted.

Drama is a form of pretence; but when the curtain comes down, the actors cease to act "in character" (we hope). The pretence is restricted.

Using theories instrumentally is a form of pretence. But we do not use these instruments without restriction: we restrict their use to those circumstances in which we expect them to work.

Suppose, though, we lift the restrictions upon these pretences. Suppose that the daydreamer starts using the "idle speculations" of the daydream as a guide to action. He dreamt that he was Napoleon, and when he "awakens" from the daydream, he acts as if he is Napoleon. The liar uses the lie to guide his own actions. The actor does not cease to act in character when the curtain comes down. The instrumental theories are used in all circumstances, with the expectation that they will work.

If these events started happening, I think that we would say: the daydreamer (now dreaming no longer) believes that he is Napoleon, the liar believes the lie, the actor believes that he is Othello (or whoever the character is), and the instrumental theory has become a belief.

Lifting the restrictions upon pretences makes them more like beliefs. Lifting all the restrictions makes them genuine beliefs (if there is no restriction upon the pretence, there is no other thinking that one does by which one could know that the pretence is "only a pretence and not what I really

believe"). Since one may lift the restrictions by degrees, pretences may become more belief-like by degrees. Since one may impose the restrictions by degrees also, beliefs may become by degrees more like pretences.

The outcome of this is that belief is more of a pretence than is "mere pretence" - for it is less restricted. It is not mere pretence: it is full-blown, unrestricted pretence.

Unrestricted pretence matches Williams' criteria for belief. It is assessable with regard to truth and falsehood.

> If a man recognises that what he has been believing is false, he thereby abandons the belief he had. (p137)

If one recognises that what one has been pretending is false, that is a restriction upon the pretence. It ceases to be unrestricted pretence, so it ceases to be belief.

> to say: 'I believe that p' itself carries, in general, a claim that p is true. (p137)

To say 'I pretend that p' carries no such claim. But someone who pretends without restriction does not say 'I pretend that p', for that would imply that he merely pretends, i.e. the pretence is restricted, and if the pretence is restricted then

he does not believe. The proper way to express unrestricted pretence is to say, 'I believe that p' (and so, 'I do not merely pretend that p').

Full-blown pretence matches Williams' criteria for belief: and so (according to Williams' account) it "aims at truth".

Yet my intuition, shared by Williams and others, is that while one can pretend virtually anything one chooses, at will, one cannot believe what one chooses, or at least not so easily. Perhaps this is because unrestricted pretences are much harder to construct than restricted pretences. They are tied in to so many more things. Full-blown pretences must fulfil some or all of these functions:

- we use them to interpret data
- we use them to guide actions
- they influence our moods and emotions
- they connect up with other items of mentation - beliefs, wishes, hopes and fears, and so on.

Performing all these functions locks a belief rigidly into place. So beliefs which perform fewer functions will be easier to alter. To alter a belief which is guiding action is like altering a gear while it is turning in a machine: even if we do not lose several fingers, the machine is going to grind to a

halt. If the machine has already ground to a halt because the gear is broken, however, we may need to replace the gear; likewise if our action has ground to a halt because our belief did not work, then we are in doubt and, to engage in practical action, we need to find another belief, or at least something that will do the job - such as using a theory, instrumentally, for limited purposes. If we replace a belief we will also (eventually) remove its influence upon our moods and emotions, and alter the connections with other beliefs, wishes, hopes, etc.

Some pretences cannot perform all these functions. For example we cannot use a contradiction to guide action - for what guidance could it possibly give us? We might use each conjunct in turn - but then each one is restricted, and so is not a "full-blown" belief.

Beliefs are like crystals: they are fixed and rigid, but they can be dissolved, "corroded by doubt". Disconnected from the role of belief they become more like "mere" pretences, fluid and shifting; so we may dissolve our beliefs, amend them in their "pretence-like" state, and then re-crystallise them. But the new crystals may be different from the old. In the pretence-like state, the belief is not there to prevent some other belief (even its own negation) being formed.

Winters [1979] says that,

> Certain mental events can occur as the immediate result of an act of will. If I wish to imagine snow falling or that I am in the Bahamas, I can do it directly; I rarely need to get myself to imagine things by an indirect route. In typical cases, imagining may therefore be regarded as a "basic" or "primitive" action. Other mental phenomena are less amenable to the proddings of the will; most notoriously, acquiring beliefs has been held to be something that one cannot do directly. (p243)

Beliefs are fixed, pretences are not. If we fix our pretences in the right way, they become beliefs. And if we unfix our beliefs, they become (mere) pretences.

Williams argues that evidence fixes beliefs, and that is why a self-deceiver must use special strategies to "steer around the evidence". I have suggested that we construct evidence by interpreting data, so that if evidence fixes beliefs then beliefs are only as fixed as our interpretations - which we may change.

Negligence theorists argue that self-deception is possible because the self-deceiver does not follow the correct procedures for enquiry. The implication is that beliefs are

justified by the procedures, whereas self-deception is unjustified. However, since the self-deceiver's beliefs are presumably as fixed as anyone else's (since they are beliefs), connection to the correct procedure is not what fixes beliefs.

We can recognise what fixes beliefs by considering how they become unfixed - how they become subject to doubt. One way is that when we are using them as instruments, they fail to work: Henry Navigator's theory of navigation causes him to steer onto the rocks, for example. If they fail to work, they cannot be true (whereas they may be false and still work, in some circumstances). If the hammer breaks, you cannot go on knocking in nails: you have to find something else to do the job. The same is true when we use beliefs as tools, e.g. to guide action. We test our theories by trying to break them - preferably before we believe them.

There is another way of unfixing beliefs. Descartes [1968] tells us the procedure:

> finding no company to disturb me, and having, fortunately, no cares or passions to disturb me, I spent the whole day shut up in a room heated by an enclosed stove, where I had complete leisure to meditate on my own thoughts. (p35)

Descartes isolates himself, as far as possible, from any need for practical action (presumably he may have put another log in the stove occasionally). He is not "disturbed by passions". He gives himself opportunity to doubt.

Pretences become more like beliefs as we fix them. There can be degrees of fixity.

> Through all the nine years which followed I did nothing but wander here and there in the world, trying to be spectator rather than actor in all the comedies which were being played there (p49 - 50)

> so, also, in order that I might not remain irresolute in my actions during the time that my reason would oblige me to be so in my judgements ... I formed a provisional moral code (p45)

In other words, Descartes put pretences in place of beliefs. The beliefs became unfixed, the pretences became (provisionally) fixed. But we might equally well say that Descartes decided to go on believing in practice what he questioned intellectually. Actions fix beliefs - they make pretences more belief-like. We may call these provisional beliefs - but most of our beliefs are provisional to some extent. When our beliefs are not locked in to actions, they

become less belief-like, less fixed. Under hypnosis ("deep relaxation") the beliefs are more completely detached from the need for practical action: they can be changed: we become more "suggestible", more open to suggestion. After hypnosis, when we act again, the suggestions become fixed again. We need fixed beliefs in order to avoid being "irresolute". When we can afford to be irresolute (especially, when we are not engaged in practical action), the beliefs can become unfixed. As spectators watching a comedy we can "suspend disbelief" (and suspend belief). As actors in a stage comedy we must pretend, but this pretence is restricted; it is not full-blown pretence, i.e. belief. Lift the restrictions, and instead of a stage there is the world, instead of a stage comedy there are our lives, and instead of restricted pretence there is (in all probability) belief.

Volition Is No "Magic Button"
(So Belief At Will Is Not Impossible)

Williams [1973] and others argue that belief at will is impossible. I agree with them that there is something peculiar about the idea of belief at will. But this is not because there is something special about belief. It is because of the way "will" is assumed to work.

Williams does not spell out his understanding of volition. At the risk of doing him an injustice, I will do so on his behalf.

There are some things that we can do at will and some that we cannot. So, for example, I can raise my arm at will, "directly", without needing to use something else to do it, whereas I cannot levitate two miles up in the air at will. If I want to go two miles up in the air then I have to find some means to do it - such as going up in an aeroplane.

Williams says - rightly - that for Hume, it is just a contingent matter of fact that one cannot believe at will. Williams argues that it is not contingent. The impossibility of belief at will is built into the grammar of the word 'believe'.

It seems to me that this makes volition into a sort of "magic button" which is located within people - only they can press it. By pressing the magic button, that is to say by "willing", they can do all the things which can be done "at will".

I object to this idea of the magic button. Agreed, it happens to be (contingently) true that I can at present raise my arm at will. But if I become paralysed then raising my arm ceases to be something I can do at will. This is because the "magic button" is connected up to a very complicated process by means of which the magic is carried out, and the process can go wrong. The fact that we only perceive the button does not give us reason to deny the existence of the process. The magic can go wrong, and this is evidence enough that there is more involved in volition than we know about.

Behind the "magic button" of the will there are processes by which the magic works. Sometimes we know what some of the processes are, other times we do not. In the case of getting oneself two miles up in the air, we know what a lot of the processes are: for instance one has to find a pilot, and an aeroplane, and probably a considerable sum of money to pay for all this, arrange a time and place when the event is to happen, travel to the airfield, etc. In the case of raising an arm we know a lot less about the processes, so we say that we can do it "directly", "at will". The upshot is that if we know about

how we do something, we cannot do it at will! This is a peculiar way to talk. It would be less confusing to talk about the things we can do voluntarily - which includes both being able to raise an arm (in some circumstances, and by mysterious processes of which we are not directly aware), and being able to go up in the air (in some circumstances, and by processes of which we are more aware).

Someone who is paralysed cannot raise an arm "at will". However if we could construct some machinery to simulate the nerve and muscle processes which take place when one raises an arm, and if we could connect this machinery up to the paralysed person's brain, then he may be able to raise an arm again, "at will", because the machinery will simulate the process which was damaged and thus prevented the "magic button" from working. If we can define the processes taking place when someone raises an arm, it may also be possible to define the processes which lead to us having beliefs. This may enable us to establish whether or not we can perform them at will.

If we believe that there are any things we do "directly" at will then we believe in magic buttons. There are always processes by which the "magic" can be explained, and the explanations always show that the things we think we do directly are indirect: they are performed by means of a process, whether we know it or not.

This means that beliefs too are produced by processes. At least one such process is voluntary, namely enquiry. Enquiry is an indirect way of voluntarily acquiring beliefs. But to call it "indirect" is misleading: it implies that there is some more direct way of acquiring beliefs. I am not convinced that "direct" is well-defined enough to say what is direct and what is noy, but in any case the supposition that there is a "direct" way to acquire beliefs is what I rather rudely called a belief in magic buttons; and the appeal to magic buttons does not help us to understand what processes are connected up to them. If we do not believe in magic buttons then we must say that every means by which we acquire beliefs is indirect.

Williams writes that in acquiring false beliefs, the self-deceiver must proceed by indirect means. This is unsurprising, since every means of acquiring beliefs is indirect. But Williams takes it to show that only Avoidance theories can provide a satisfactory account of self-deception:

> that ... is the project of the man who is deceiving himself, and he must really know what is true; for if he did not know what was true, he would not be able to steer around the contrary and conflicting evidence (Williams [1973], p.151, my emphasis).

Williams' argument is sophisticated. But when it is put crudely, I think it amounts to this: the will is not the magic button which can create beliefs - but evidence is!

In the preceding chapters I presented arguments to show that:

- evidence is not a "magic button" for the production of beliefs any more than is "the will"
- therefore the self-deceiver does not need to "steer around" (i.e. avoid) "the" evidence, since "the evidence" does not compel belief.

My aim now is to counter the arguments which Williams uses to show that belief at will is impossible. My main counter-argument is that there are processes by which we acquire beliefs, and that these processes can be voluntary. Later on I shall use this argument to defend Role Simulation theories.

Williams argues that belief at will is not possible because "belief aims at truth". This is an odd expression to use, and it seems to invite paradox. We have aims, but beliefs do not. Even if in some sense belief "aims" at truth, it would not follow that people acquire beliefs by aiming to acquire truths. However this is, I think, what Williams needs for his argument to succeed. Williams explains what he means by "belief aims at truth" in the following way:

1. "truth and falsehood are a dimension of an assessment of beliefs as opposed to many other psychological states or dispositions". (p137)

2. "To believe that p is to believe that p is true"; "if a man recognises that what he has been believing is false, he thereby abandons the belief he had." (p137)

3. "To say 'I believe that p' itself carries, in general, a claim that p is true." (p137)

Williams then argues as follows:

> If I could acquire a belief at will, I could acquire it whether it was true or not; moreover I would know that I could acquire it whether it was true or not. If in full consciousness I could will to acquire a 'belief' irrespective of its truth, it is unclear that before the event I could seriously think of it as a belief, i.e. as something purporting to represent reality. At the very least, there must be a restriction on what is the case after the event; since I could not then, in full consciousness, regard this as a belief of mine, i.e. something I take to be true, and also know that I acquired it at will. With regard to no belief could I know - or, if all this is done in full consciousness, even suspect -

> that I had acquired it at will. But if I can acquire beliefs at will, I must know that I am able to do this; and could I know that I was capable of this feat if with regard to every feat of this kind which I had performed I necessarily had to believe that it had not taken place? (p148)

I disagree with Williams' argument. He writes that, "it is unclear that before the event I could seriously think of it as a belief, i.e. as something purporting to represent reality." Being unclear is not the same as being impossible; and I suggest that it is not impossible. Suppose that someone else believes it "before the event", i.e. before I believe it. Then I have no trouble in thinking of it as a belief: it is not a belief of mine (as yet), but it is someone's belief and so, in Williams' words, "something purporting to represent reality". If I am able to become sufficiently like them, I will be capable of believing it too.

I think that what Williams wants to say is that the belief I already have prevents me from believing something else inconsistent with it (such as its negation). What the self-deceiver needs, then, is some way of suspending belief and disbelief. "Suspending disbelief" is an expression which is often applied to what we do when we read a novel or watch a play. It suggests a way to explain the manner in which

beliefs can be changed: by suspending belief and disbelief we enable ourselves to use imagination to construct new beliefs. That process of construction may alter the ways we think, act and feel, so that when we stop suspending belief and disbelief, our beliefs are different to what they were before. The process might be like the crystallisation and dissolution of salts: fixed and frozen beliefs become fluid and thawed, they are altered, and then crystallised again into fixed and frozen beliefs. Maybe we always have beliefs (stored away somewhere in the retentive material of the memory), but we are not always believing them - using them, that is. Sometimes one forgets something one believes, and cannot recall something one knows.

Armed with this metaphor, we need not suppose that a self-deceiver somehow "masks" one belief in order to believe something else which is inconsistent with it. He need not mask the belief because he need not have the belief. It may be stored away in memory along with all the other things he once believed and believes no longer; but he does not need to "mask" the belief, since he can destroy it outright.

This approach has the added advantage that we need not discuss beliefs in isolation from all the other things we do - imagining, wishing, feeling, wondering, hoping and fearing, surmising, hypothesising, and so on. Role theories of self-

deception use this approach. I discuss Role theories later on. Before doing so I lay the groundwork for that discussion.

We know there must be some process by which beliefs change: otherwise we would all be in the position of the bigot whose staunch beliefs immunise him from any other way of thinking. That seems a caricature of what normally goes on when someone believes something.

Williams implies that evidence is what changes beliefs: evidence can compel us to believe what formerly we did not believe, or to disbelieve what we formerly believed. "If I could acquire a belief at will, I could acquire it whether it was true or not. Moreover I would know that I could acquire it whether it was true or not." But we do acquire beliefs whether they are true or not, often despite our best efforts to believe only what is true. We do acquire them through a voluntary process - enquiry, for example - and although this process does not occur "just like that" (as Williams puts it), it is not in that respect different from any other process by which beliefs are generated: there is no direct method, no magic button for acquiring beliefs. When we do enquiry, we know that we are aiming to voluntarily acquire true beliefs, and, if we arrive at a belief, then we know that we believe it, whether or not it is true. We hope that it is true, and since we believe it we at least do not believe that it is false. But unless we think

that we are infallible then we will usually concede that we may be wrong: and in that sense, we know that we have acquired it whether it is true or not, and have done so voluntarily.

Williams' argument seems irrelevant to the question of whether or not we can believe something at will. If his remark is intended as a counter-factual conditional and therefore as a disproof of the claim that we can acquire beliefs at will, then I think it is a failure; for the consequent, 'I could acquire it whether it was true or not' is true. We do acquire both true beliefs and false beliefs, and we know that we do. Only we do not know which are true and which are false. Let's look at this another way, running an argument in parallel with Williams' argument:

- If I can acquire a belief against my will, then I can acquire it whether it is true or not. Moreover I may know that I can acquire it whether it is true or not. So in full consciousness I could acquire a 'belief' against my will irrespective of its truth: could I seriously think of it as a belief?

The answer is, "yes, I could". Think of someone who knows he is about to be brainwashed, against his will. He may know that the brainwashing will be successful, regardless of whether or not the beliefs he acquires as a result are true or not. Now

consider someone who is going to be voluntarily brainwashed. He too may know that the brainwashing will be successful, whether or not the resulting beliefs are true.

Williams concedes that this sort of case is possible: one may use drugs or use the services of a hypnotist in order to acquire false beliefs; but although it is possible, it is "very deeply irrational", and it is not "directly" at will.

I have argued enough about the "indirectness" of all belief-generating processes, voluntary or not. Williams' admission allows us to undermine the whole argument. For Williams assumes that hypnotism and the use of drugs are wholly different from the normal means of acquiring beliefs (whatever they are); but one can argue that the use of drugs and hypnotism and (for good measure) brainwashing may all work because they utilise the normal means of acquiring beliefs. Unless Williams can show that normal belief-acquisition is relevantly different from these "abnormal" means, then his argument will reduce to the normative and not widely disputed suggestion that we ought to aim for the truth.

Belief against one's will seems as puzzling as belief at will, when we put it into the form of Williams' argument. It is irrelevant that we can acquire beliefs which are true and other beliefs which are false. If you acquire the belief then you

cease to believe that it is false (if you ever did so). Better to ask how someone comes to believe something which they previously thought was false.

The answer I reconstruct from Williams' article (i.e. the answer which I suppose Williams would give) is that such a change of belief could come about when someone encounters fresh evidence: "in saying that his belief is based on particular evidence, we would mean not just that he has the belief and can defend it with the evidence, but that he has the belief just because he has the evidence. This says that if he ceased to believe the evidence then, other things being equal, he would cease to have the belief."

I think that there is something odd about this argument too. As a counter-example, lets imagine the following anecdote, recounted by a character whom I shall call Henry Navigator:

"I formed a hypothesis about the relation between the earth and the fixed stars; I leapt to the conclusion that the hypothesis was true, and I used it to make decisions about how to navigate my ship. The hypothesis worked well for this purpose: I always managed to navigate to the places I wanted to visit. I was not forced by compelling evidence to believe the hypothesis. I leapt to the conclusion: in other words, I

believed at will. Fortunately, the hypothesis turns out to be true."

Henry Navigator claims that he can believe the hypothesis at will - for he thinks he did it, by "leaping to the conclusion". He does not claim to know that the hypothesis is true; he can even concede that it is highly probable that the hypothesis is false, and still say truthfully, "nonetheless I believe it". "After the event" he is capable of regarding the hypothesis as something he takes to be true and also knows that he acquired it at will. "Fortunately, the hypothesis turns out to be true".

He would claim that the belief is true, wouldn't he, since he believes it. So if the example is described correctly then it is possible to believe at will, and to know that one's belief was acquired at will; at the time he made the "leap", Henry Navigator could also know that he was sustaining his belief at will, unsupported by evidence.

What is more difficult, indeed impossible, I think, is to acquire a belief at will and then, after the event, both sustain the belief and believe that it is false. When (or if) Henry Navigator stops believing his hypothesis, he will still be capable of knowing that he believed it at will, and he will also be capable of knowing that he believed at will something

which is false. So Henry knows that he is able to believe things at will. He also knows that the things he believes at will may be false - he sometimes makes mistakes. However, he thinks that all the things which he presently believes at will are true, while admitting that they may be false. He believes them: he does not claim to infallibly know all of them.

Henry Navigator might also tell us that: "when I invented the hypothesis I was just playing with ideas: I invented the hypothesis not with the intention of gaining a truth, but for the sake of an amusing and interesting fiction which, I believed, was false. But then it occurred to me that it would be a terrific help to navigation if it were true, so much so that I decided it was worth trying it out: suddenly I had a hunch that it was true." Henry moves from the belief that the hypothesis is false, towards "trying it out" - a stage which we could call tentative or provisional belief - until the hypothesis has become an habitual belief. At this stage he still does not have evidence of any "compelling" sort that the hypothesis is true: he only has evidence that the hypothesis works for his purposes - and false hypotheses may work as well as true ones, for limited purposes.

Believing a hypothesis is somewhat like trusting a person: one may trust someone more after long experience of their character and conduct has shown them to be worthy of trust; but

sometimes one may trust as an act of will, knowing that there is no basis of evidence or experience to show that the person is worthy of trust.

I claim that Henry Navigator has provided us with a counter-example to Williams' claim, and I do not detect the difficulties which Williams claims there must be. Suppose that Henry Navigator finds evidence which contradicts his belief about the relation between the fixed stars and the earth. He is entitled to reject one or more of the premises which led to the contradiction. He genuinely believes the hypothesis: so he rejects the evidence. On the basis of his belief he "knows" that the "evidence" must be false.

So how is the evidence supposed to change Henry's belief? Henry must first, as it were, put the belief in suspense: for if the belief is being used, then it ensures the rejection of the evidence. Henry must put the belief in suspense before the evidence can have any effect: the evidence cannot do it for him. Therefore evidence alone cannot alter beliefs. Evidence against Henry's hypothesis, for example, cannot alter his belief in the hypothesis unless he does something to allow it: and this may well be something that he does at will, namely what I called "putting the belief in suspense". He may have strong motives for doing so: for example, he may encounter the evidence because his navigation does not lead him

to the places he wants to visit: the hypothesis breaks down, it does not "work": Henry is then strongly motivated to throw out the hypothesis, as he would any other broken or useless tool - or else find ways to mend it. To do so he will probably want to study the conditions which broke the tool - and these conditions are "the evidence".

Williams thinks that there are ways of acquiring beliefs other than by the action of "evidence":

> not every belief that I have which is based, is based on evidence. There are some beliefs that I have which are not (relative to the probability of their being true) random or arbitrary, and which are very proper beliefs to have, but which are not based on other beliefs that I have. Indeed, there is a very good reason why it cannot be the case that every belief which one has is based upon another belief one has - namely, that one could never stop (or start). Quite evidently there are non-random beliefs which are not based upon further evidence. The most notable of these, of course, are perceptual beliefs, beliefs that I gain by using my senses around the environment. (p143)

If this is the case, then I do not need evidence to support a belief. If I believe p on the basis of evidence e, and I cease

to believe e, then I can continue to believe p provided that I treat it as a non-random belief. For who is to say what we are entitled to classify as a "non-random belief"?

I deny Williams' claim that "one could never stop (or start)" unless one had beliefs which are not random or arbitrary (relative to the probability of their being true). We may start from any arbitrary hypothesis. We run the risk of being wrong, of course, but we may not have the option to start from anything else. We are forced to start from whatever circumstances we are in.

Suppose that we lack the non-random beliefs with which Williams thoughtfully provides us. We would not become helpless. We could adopt a hypothesis at random, and try it out in the hope that it works. We are frequently cautioned against this: it is "leaping to conclusions", or something very like it. Leaping to conclusions is justified, provided that we are prepared to test the conclusions and be proved wrong. It is a justifiable way to proceed because there is no other way to get started. It is the way we still "get started" in areas of enquiry which are unfamiliar to us (and everyone else), e.g. some parts of scientific enquiry.

As for the beliefs we "leap" to, they are justified (if at all) by standing up to testing. They are not justified in advance

of testing by our giving them a special "foundational" status. Our perceptions are refutable, and often are refuted by our theories - theories which may well be true.

Williams argues that "perceptual beliefs" are not based on further evidence. It is strange, in that case, that people's perceptions differ according to the culture within which they grow up.

Williams' claim that "one could never stop (or start)" is coloured by his foundationalism; I do not think that it is the "evidence" which led to the foundationalism. For unless one is inclined to foundationalism to begin with, the argument does not seem very convincing. One can "start" from any random hypothesis: if it is falsified then you will know that it is false; if it is never falsified then at least you will not suffer the awful consequences of believing something which does not work (a theory which leads you to navigate onto the rocks in a raging sea, for example); and if it is false but not falsified then you will continue to believe something false, unless you can invent a better hypothesis - and so would anyone else, including those people who claim that they have non-random beliefs which are foundational.

Williams' claim that "belief at will is impossible" presupposes the truth of foundationalism. For belief at will provides an

alternative to foundational beliefs. Foundationalism is false. We do not need foundational beliefs to "get started". Even hypothesising at will provides an alternative to foundational beliefs.

When someone adopts a hypothesis as a guide to action, the hypothesis gains a more belief-like role in his or her thinking. If it becomes the person's habitual guide to action, then it looks even more like a belief. Ask Henry Navigator if his hypothesis is true: "I believe so", he replies. Ask him why he believes it. "I've tried it, and it works". But falsehoods can work too, sometimes, can't they? "Yes, I did not say that I know it, only that I believe it: it could still be false. But do you know a better way of finding out?"

If belief at will is possible, then Williams' non-arbitrary foundational beliefs are not necessary: one can get started without them. Admittedly the beliefs do not get justified without being put to the test. But why suppose that our beliefs can (or should) be justified before they are put to the test?

Williams claims that there are non-arbitrary foundational beliefs. Are they a better way of finding out than Henry Navigator's arbitrary hypotheses? Perhaps not: for although they are (allegedly) foundational, that does not guarantee that

they are true. Perhaps they are ineradicable falsehoods. Perhaps they are eradicable falsehoods. If they are foundational, they cannot be corroborated or refuted by evidence. Yet evidence can support or falsify "perceptual beliefs", for example. If "perceptual beliefs" are foundational, how is it that perceptual beliefs can be tested in practice: when the perceptual belief passes the test, that is evidence for the truth of the perceptual belief: and if it does not pass the test, then it is falsified. So, to summarise:

- foundational beliefs do not exist
- but there must be some way to get started
- we get started by belief at will
- if belief at will is impossible, then we cannot get started
- so belief at will must be possible.

This argument is a mirror to Williams' argument for foundational beliefs, which goes like this:

- belief at will is impossible

- but there must be some way to get started

- we get started by acquiring foundational beliefs

- if there are no foundational beliefs, then we cannot get started

- so there must be some foundational beliefs.

I claim that we can believe at will, not in the sense of willing "directly" - using the magic button - but in the sense that sometimes we can voluntarily alter the normal processes by which beliefs are acquired. One of the voluntary processes for acquiring beliefs is "leaping to conclusions"; and, in an argument which paralleled Williams' argument for foundationalism, I argued that "leaping" is the only way our beliefs get started. Therefore "leaping" is a justifiable strategy. But it does not justify the beliefs to which we leap. That justification comes, if at all, when we test our beliefs.

So far as I know, nobody doubts that we are able to leap to conclusions. Many people doubt that there are foundational beliefs. So if we assessed our arguments on an electoral basis, more people would vote for my claim than would vote for the claim Williams makes!

Against my claim it can be argued that someone who leaps to conclusions does so on the basis of evidence - inadequate evidence, to be sure, but evidence nonetheless; therefore my argument does not show that there must be beliefs at will.

I agree that we sometimes leap to conclusions on the basis of evidence. That does not show that there is always evidence when someone leaps to conclusions: "inadequate evidence" can mean: "no evidence at all".

There is an equivocation on the word 'evidence' which could make my argument seem less appealing. In one of the preceding chapters I distinguished between "evidence" and "data", and argued that:

1. evidence can "compel" belief but only if we construct it - and the construction of evidence can be voluntary

2. we do not construct data, but nor does it "compel" belief.

The idea that evidence is a "magic button" (which compels belief) is due to using the word "evidence" to refer to both what I call data and what I call evidence. Since it is up to us whether or not to construct evidence, evidence can only "compel" our beliefs if we consent to be "compelled". However, there may be plenty of things which can coerce us into giving

that consent, or witholding it. These forms of coercion are more to do with our motives than with data or evidence.

My argument against Williams may provoke incredulity: "do you mean to say that we can believe whatever we want!"

That is not quite what I mean (though it is supported by, for example, Julius Caesar: "men willingly believe what they wish" - Caesar [1951], I.iii.18). I agree with Hume, that sometimes belief responds to the will, sometimes it does not. But this "merely a contingent fact" can be explained by describing the processes by which beliefs are acquired.

I am not suggesting that we can always believe whatever we want. Sometimes we can believe at will things that we do not want to believe. For example, one may not want to believe that one is less than perfect: it may be much more comfortable to believe one is flawless. But one may also want to believe the truth, and the truth may be that one is not flawless. The desire for truth may be stronger than the desire for comfort.

Let's take another example: Henry Navigator "leaps to the conclusion" that his hypothesis is true. I do not see that "wanting" enters into it (it could do, of course, if we set up the example in the right way; and I shall suggest some examples later). Henry believes the hypothesis "at will" but

it would be misleading to say that he believes it because he wants to: that would suggest desires which Henry may not have. In particular, it may suggest that Henry wants to believe the hypothesis whether or not it works - which would be false.

Williams argues against belief "directly" at will, conceding: "but there is room for the application of decision to believe by more roundabout routes": "for we all know that there are causal factors, unconnected with truth, which can produce belief: hypnotism, drugs, all sorts of things could bring it about that I believe that p" (p149). Someone could use these factors to acquire a belief independent of considerations of truth. Williams add that such a project is:

> very deeply irrational, and I think that most of us would have a very strong impulse against engaging in a project of this kind however uncomfortable these truths were which we were having to live with. (p150)

My reply is: perhaps, perhaps not. I do not know what impulse most of us would have. It is misleading to talk about "roundabout routes" for it implies that there is some more direct route, when there is not. For Williams this more direct route is provided by a magic button, namely evidence.

Williams concedes that rationality is a desideratum, a norm rather than a psychological law. He also concedes that there may be causal factors unconnected with truth which can produce belief. These factors seem to me worth considering: we might find, for example, that the elements of hypnosis which make it an effective generator of belief can also be used "directly", at will.

There might be many more of these "causal factors unconnected with truth": and some of them might be the starting points which Williams claims we do not have unless we have foundational beliefs which are not arbitrary. Causal factors unconnected to truth can give us starting points for belief; they might not be justifiable starting points in Williams' sense of not being arbitrary "relative to the probability of being true"; but I see no reason to suppose that our starting-points are justifiable. Justification by results - because they work - or by logic - because they are logical truths - are the only kinds of justification I will concede: there is no reason to suppose that there is any sort of prior-to-experience justification of the kind needed for foundational beliefs.

There is also disagreement about what constitutes "rationality". There might be a trade-off between considerations of truth and rewards one might gain from disregarding truth considerations.

Consider some cases in which beliefs are generated by causes "unconnected with truth". For example, consider brainwashing. Why does it work? Here is a suggestion: brainwashing is achieved by making it extremely unpleasant for the victim to express his own beliefs: expressing those beliefs meets an extremely hostile reception, including verbal and perhaps physical abuse. This leads the victim to outwardly conform to the demands of the people doing the brainwashing.

Outward conformity requires the victim to do a certain amount of information processing: he needs to be able to predict what response is required, and make that response. The brainwashing sessions crowd out any time the victim might have had to think his own thoughts. The victim's own beliefs are never rehearsed, never aired, never used: they fall into decay and gradually the responses he is compelled to make create a changed psychic environment or "mind set": outward conformity leads to inward conformity too.

The information processing which he is compelled to perform then takes on the role of beliefs and the former beliefs cease to perform that role: they join the store of unused hypotheses, speculations, etc, which everyone carries around in their heads. In short, the victim's beliefs fade away from lack of use, while he is forced to use the brainwashers' proposals as though they were beliefs - and thereby they become

beliefs. He "gets the habit" of acting as though he believed them. But this habit may become indistinguishable from belief.

If brainwashing works in the way I am suggesting, then it is only an extreme case of the "socialisation" we all go through by virtue of living in society with other people. If brainwashing works then so (in a less extreme but no less all-pervading way) will socialisation. The sort of socialisation will be relative to the sort of society and to one's position within it. If it is possible to change one's position in society then it is also possible to adjust the kind of socialisation to which one is subjected.

Let us take a further instance of causal factors influencing beliefs in a way unconnected with truth: subliminal advertising. Suppose that a cinema audience is subjected to a momentary message flashed on the screen; the event is over so quickly that the audience is not aware of the message. The message says "buy icecream!" and, in the interval, the cinema sells more icecream as a result.

What happens now if a member of the audience is asked to explain why he bought icecream in the interval. Probably he gives an explanation in terms of beliefs and desires. We, however, are in the know: we know that there is a well-tested observation that subliminal advertising can lead people to buy

icecream; and we know that this member of the audience has been subjected to subliminal advertising. We also know that similar sorts of explanation are forthcoming when people are asked to explain actions which they performed in response to post-hypnotic suggestion. For example, someone under hypnosis is given a suggestion that after the hypnosis session is over, he will open a window when he sees the hypnotist make a signal; the session comes to an end; after a few minutes the hypnotist gives the signal, and the person who was hypnotised goes and opens the window. During those few minutes he is also observed to be covertly watching for the hypnotist to give the signal.

He is then asked why he opened the window, and gives an explanation: for example, that the room felt stuffy and he needed some fresh air.

What has happened? The person who was hypnotised seems unaware that he responded to a post-hypnotic suggestion, and also unaware of watching for the signal - even though observers could see him "keeping an eye on" the hypnotist. If we are correct in saying that the action of opening the window resulted from the suggestion made while he was under hypnosis, then where did the person's own explanation of the action come from? He must have invented it, and by "invented it" I do not mean a deliberate lie: I suggest that the explanation was inferred from the behaviour, in just the same way as an

observer might make the same inference with the same limited information: for example, the observer might have come into the room after the session of hypnosis and before the signal was given.

If my suggestion is correct then we have found a case in which someone infers his own beliefs and desires from his own behaviour, i.e. was deceived by his own behaviour in just the same way that someone could be deceived by the behaviour of another person. For the behaviour to be deceptive we must make the inference, and to make the inference we must presuppose that if there is an action then there will be beliefs and desires which explain it. But in the case we are considering it seems that the beliefs and desires do not really explain the action, indeed the beliefs and desires may never have occurred (for example, we can ensure that the room is not stuffy, so that there is no basis for the person to believe that it is stuffy or to wish for fresh air).

It is hard to imagine an action that cannot be explained by reference to beliefs and desires, for we are so inclined to think that there must be some such explanation if only we can find it; in cases where we cannot find it, we may still feel that there must be an explanation, or we may decide that the "action" was not an action after all but something which merely happened to the person, like a knee-jerk reaction.

So how does subliminal advertising or post-hypnotic suggestion work? Not through the ratiocinative processes of critical thought, not through the processes of reason-giving and justification in terms of beliefs and desires): those processes are bypassed. Some other process or processes operates, and it seems inappropriate to apply the explanatory apparatus of "beliefs and desires" to it - for when we do try to apply that explanatory apparatus the result is that we make misleading inferences and draw false conclusions.

Suppose that someone does make the inference to explain his own behaviour, e.g. "the room is stuffy", to explain the action of opening the window. Although the inference is false, nonetheless it now becomes his belief that the room is (or was, prior to opening the window) stuffy: the inference from the action generates the belief: it generates "evidence" which seems to confirm its correctness.

Now consider someone who does something "for no reason at all", "out of high spirits", etc: let say that Henry has too much to drink, sees a policeman and knocks the policeman's hat off - for no reason at all, "for its own sake", "for the fun of it". Summoned to court to explain his action, Henry says that the policeman was harassing him, looked at him in a very aggressive way, and so on. And Henry seems to sincerely believe what he is saying. Perhaps he really is sincere - as sincere as the

person carrying out the post-hypnotic suggestion. He has been asked to explain his behaviour and he gives the best explanation he can. It just happens to be false.

We can only detect someone's beliefs and desires through their behaviour - the things they do and say. Sometimes they too may make the inferences from their own behaviour: they are then in no better position than us to decide what beliefs and desires (if any) may have been involved in the behaviour. Someone might perform an action for its own sake, knowing that by doing so he was committed to having the action explained in terms of beliefs and desires; and yet, prior to the action, he never experienced such beliefs and desires. The only reason for positing such beliefs and desires is that they are required for explanation of the action: the action is taken as a prototypical action for those beliefs and desires: and so the beliefs and desires are, as it were, created by "back-projection" from the action.

Someone who wanted to have those beliefs and desires, then, could gain them by means of the behaviour: the behaviour would create them by "back-projection". Pascal suggests something like this for those who wish to gain religious faith: "go to mass and take holy water" is his advice: behave like a believer, and the beliefs will follow.

> we are as much automaton as mind. ... we must resort to habit once the mind has seen where the truth lies, in order to steep and stain ourselves in that belief which constantly eludes us, for it is too much trouble to have the proofs always present before us. We must acquire an easier belief, which is that of habit. With no violence, art or argument it makes us believe things. (Pascal [1966], p274)

But, it might be argued, the case is different when someone behaves that way with the aim of gaining a belief. In the case of post-hypnotic suggestion, there may have been no belief or desire when the behaviour was taking place: the belief and desire posited by the explanation could therefore occupy the place which was, prior to then, a vacuum. But someone who adopts behaviour in order to acquire a belief already has a desire and belief which explain the behaviour: "back-projection" will therefore try to fill a space which is already occupied: an explanation in terms of beliefs and desires already exists, so that the back-projection will fail.

It is by no means clear that the back-projection will fail. After all, if behaviour can create beliefs in the case of brainwashing, why shouldn't it do so in this case? Why should behaviour be an effective generator of belief in one case but not in the other?

"The Mister Men": Three Cases Of Self-Deception

Mr Negligent, Mr Mobile and Mr Radical all live in the same small community. Any news does the rounds very quickly, so they all have access to the same data; yet their beliefs differ, for they interpret the data in different ways.

Mr Negligent can claim the sanction of tradition for his way of interpreting data: it is the way anyone (that is, "anyone who is anyone") would interpret data, the "natural interpretation". What this means in practice is that if you do the sort of things Mr Negligent does in the circumstances which Mr Negligent occupies, then to interpret data as Mr Negligent does takes no noticeable effort whatever.

Mr Negligent takes it for granted that the natural interpretation must yield true beliefs. Other people in the community are less confident. Mr Radical, for example, thinks that Mr Negligent has false beliefs.

Mr Radical points out that tradition can sanction mistakes. Uncritical conformity with tradition can therefore lead us to acquire false beliefs.

Sometimes we can detect and correct our false beliefs simply by thinking about what we already know, without having to seek any further evidence. Mr Negligent could do so, for example. But, in Mr Radical's view, it suits Mr Negligent to have the beliefs he has. For while it is true that having false beliefs can have very deleterious consequences, in Mr Negligent's case the awful consequences befall other people, while Mr Negligent reaps rewards. Mr Negligent does not mind this since he lacks sympathy for the suffering of others and shrugs off his responsibility for causing those sufferings; the social setup is such that he can get away with it.

Mr Radical holds Mr Negligent responsible for having false beliefs. He points out that Mr Negligent could easily have corrected his mistakes, only it suited him not to do so. Mr Radical calls Mr Negligent a self-deceiver. He blames Mr Negligent not for what he does, but for what he fails to do; not for what he intends, but for the lack of an intention which he ought to have had.

Mr Mobile is one of those who suffers as a result of Mr Negligent's false beliefs. But Mr Mobile does not blame Mr Negligent. He admires and envies Mr Negligent although he is sure that Mr Negligent's beliefs are false. It seems to Mr Mobile that it would be worth putting up with the false beliefs for the sake of the rewards reaped by Mr Negligent. For Mr

Mobile reasons that the false beliefs are a by-product of behaving as Mr Negligent does in the circumstances Mr Negligent occupies.

Mr Mobile takes Mr Negligent as an exemplar of the recipe for success. By following the recipe Mr Mobile aims to gain the same results as Mr Negligent - both the rewards and the beliefs. For the beliefs play a role in protecting Mr Negligent from any pangs of guilt for the way he is behaving.

So in Mr Mobile's view, Mr Negligent is a role-model for belief-acquisition: Mr Negligent has (or perhaps we should say is) the recipe by which one can acquire the particular beliefs which he has.

Mr Radical regards Mr Mobile too as a self-deceiver.

Mr Negligent, while content to turn a blind eye to his own shortcomings, is more than willing to analyse Mr Radical's failings.

This is what Mr Negligent has to say: any natural interpretation of the available data would lead to the conclusions which I have come to myself, but of course these conclusions do not suit Mr Radical. He cannot find any other data - although he would like to - and no other interpretation

exists, since no interpretation other than the natural one has hitherto proved necessary. So Mr Radical replaces the natural interpretation with sheer invention. He invents a new, distorted interpretation of the data.

If only Mr Radical would follow the tried and tested reasoning which comes so naturally to all of us (sighs Mr Negligent), then he would inevitably arrive at the same conclusions as we do. For that reasoning is the process by which we generate our beliefs. But he avoids that process: and you cannot have the beliefs without having the process which generates them.

Instead of the tried and tested reasoning, Mr Radical substitutes some other process in its place. He consequently arrives at different beliefs. But what justifies this substitute process? Why, there is no justification whatever. Where did this process come from? Mr Radical invented it! What sort of basis does that give for belief? Not a very reliable one, suggests Mr Negligent. What is Mr Radical, but a radical self-deceiver!

Mr Radical, though, has his reply to Mr Negligent: where does Mr Negligent's traditional interpretation come from? It too was invented - we inherit not only the wisdom of the ancients but all their entrenched folly also, as Nietzsche pointed out. The traditional interpretation has been tried and tested - and

found to have appalling consequences, though it suits Mr Negligent not to notice them. The justification for the Radical interpretation must be through its results - trying and testing is the only way to justify any interpretation. Mr Negligent is not entitled to complain that it lacks justification, when it has never been put to the test. Mr Negligent would rather that it never was tried.

Lets put these three characters into a situation and watch how they behave, and how they disagree.

Here is the situation: Mr Negligent is a senior manager working for a large company. Mr Mobile is a junior manager for the same company, and Mr Radical is a still more junior member of the staff.

Mr Negligent believes that promotion within the company is on the basis of merit, merit being the ability to do one's job well. It suits him to believe this, for if it turned out that his own promotion had something to do with being a nephew of one of the owners, for example, then his position, authority and self-esteem might all be called into question.

Mr Mobile is firmly convinced (rightly or wrongly) that promotion within the company depends upon being accepted as a member of the "old boy network" and has very little to do with

merit. However, he knows that to say so would ruin any chances of promotion which he might have. He also knows that if he stays on and outwardly conforms in a way which will gain him promotion, then that conformity will shape his attitudes and beliefs as well. In the end he will become just like old Negligent, who really believes (on very thin evidence) that promotion is on the basis of merit.

Mr Radical too is firmly convinced that promotion has very little to do with merit. He may be right or wrong in this respect, but it is important for him to believe it since he has already been passed over for promotion a number of times: his self-esteem is at stake. There is very little evidence to support his belief (apart from old Negligent being a senior manager, but that might be just a ghastly mistake that someone made a long time ago).

Having formulated the theory that there is an old boy network, he finds that it explains and predicts a lot of things, such as the way Mr Mobile seems to be progressing so rapidly. Seemingly disconnected events begin to fall into patterns. His theory seems to work.

If we create patterns of thinking, then it is not surprising that events fall into the patterns - even if the patterns are not out there in the events but only in our thinking.

So who is right? Suppose promotion is genuinely a result of merit. Mr Negligent therefore turns out to have true beliefs, but this is really a matter of epistemic luck since he has never made any attempt to test his beliefs. Mr Mobile has false beliefs, but his beliefs will become true as his circumstances mould them; of course, he deserves very little congratulation for this fortunate state of affairs since his regard for promotion far outweighs any regard he may have for the truth.

Mr Radical has achieved false beliefs, but this too is a matter of epistemic (bad) luck: they could just as easily have turned out to be true.

All three characters have adopted strategies typical of self-deception, although only one has acquired a false belief as a result. Their strategies would have been the same if the facts had been different. Suppose the truth had been that promotion was not given on merit. In that case Mr Negligent would be self-deceived, Mr Mobile would have been headed for self-deception and Mr Radical, though lacking a false belief, would have adopted the strategy of a self-deceiver.

The three strategies I have mentioned are not paradoxical. The next few chapters explain why.

Negligence Theories

The basis of Negligence theories is simple: Negligence theories recognise that it is very easy to make mistakes, and not always so easy to avoid them. The mistakes can result in our having false beliefs. To avoid the mistakes may require a degree of diligence. We are not always forced to be diligent: we may not make the effort, we may instead neglect to do things which would prevent our having false beliefs. So we may be responsible (in rather a strong sense of the word 'responsible') for being mistaken, or for being in ignorance.

Negligence is the prototypical way of being responsible for something of which one is wholly unaware, something which one never intended. For example the negligent driver may cause an accident: he did not intend to cause the accident, nor did he intend not to cause the accident: he never gave it a thought. He was not aware of causing the accident - not until it was too late to do anything about it.

Take another example: a metallurgist is testing a sample of aluminium to check that it conforms to the standard required. She is negligent: without testing the sample properly, she

confirms that the aluminium conforms to the standard. But It is sub-standard, though she does not know it. She confirms "in good faith" that the aluminium is of the quality required. As a result, the aluminium is used in building an aeroplane, and shortly afterwards the wings fall off while the aeroplane is flying over the Atlantic. She did not intend this to happen; she was wholly unaware that it was going to happen. Nonetheless she is responsible: the disaster happens because of her negligence.

Being mistaken or in ignorance may suit us very well. For having false beliefs may not be disastrous for us: sometimes we benefit while other people suffer the disaster on our behalf. If, like Mr Negligent, we lack sympathy for other people, then we may not feel moved to alter the situation which benefits us.

For example, a false belief about one's own conduct can give the benefit of a quiet conscience; an easy bigotry can sidestep the torments of self-doubt and self-condemnation, while smoothing the way to picking up the glittering prizes at someone else's expense. One can maintain a sort of sensitivity while not bringing it to bear upon one's own acts: a tyrant can watch theatrical productions and participate in the experiences offered to the audience, condemning the deeds portrayed though he does the same and worse every day.

Negligence theories show that there can be non-paradoxical cases of self-deception. People with false beliefs, who are able but unwilling to have true beliefs, can be counted as being self-deceived through negligence.

However there are, in my opinion, instances of self-deception which cannot be explained by negligence theories. Negligent self-deception is passive: it does not create false beliefs (or false doubts, or redeemable ignorance), it just neglects to dispel them. It relies on something else to generate the beliefs, something which is already in place. Butler suggests that "self-love" performs this role. Plato (as interpreted or misinterpreted by me) implies that the role is performed by the appetites or by ambition having mastery over reason, and the self-deceiver neglects to "restore the balance".

Yet there may be instances of self-deception where the deception is not derived from a pre-existing source of error. There may be cases in which the generator of the belief has to be invented, or sought, rather than just being effortlessly available.

In later chapters I go on to consider what it is that generates the deception. For negligence alone is not sufficient to explain self-deception. It describes what the self-deceiver does not do, but not what he does.

The strength of negligence theories is that they do not require that the self-deceiver knows or intends what he is doing. Negligence is a way of being responsible through lack of knowledge and intention. If we now look for the positive generators of deception, we may lose that advantage, and have to explain, in the face of the paradoxes, how the self-deceiver can know and intend the deception and still be deceived.

A weakness of some Negligence theories (e.g. Mounce [1971], Peterman [1983]) is that self-deception is treated as an aberration from the normal way of acquiring beliefs: and this normal way of acquiring beliefs is taken to be something like doing enquiry in a "proper" way. The proper way is taken to be one which conforms to some set of epistemic standards or guidelines for seeking truth and avoiding falsehood.

This treatment seems to me an inversion of the true situation. The epistemic standards are subjects for disagreement; they are developed over long periods of time, with great effort and many mistakes. By contrast, self-deception has no methodology, no historical development, and seems often to be effortless and unerring.

Negligence can result in false beliefs only if the false beliefs are acquired by doing what comes naturally. Self-deception, therefore, can hardly be correctly described as an

aberration from the usual ways of acquiring beliefs. A better description would be that self-deception is one way of doing what comes naturally, unconstrained by epistemic norms. It is better to explain the epistemic norms as a development of "doing what comes naturally" rather than to try to explain self-deception as an aberration from the norms. For one thing, all the epistemic norms that I recall serve to restrain belief, not to generate it: that is why scepticism plays such a large role in the development of epistemology as a methodology of enquiry, and why creativity has not been given such a role.

Negligence alone seems rather a restricted strategy for self-deception. Indeed in the absence of motives of an appropriate kind (discussed below), negligence may not be self-deception at all.

For example, someone may neglect to find out just how many earwigs there are in Europe, because he is more concerned to find out where his next meal is coming from: epistemic negligence which is localised may be explained and neutralised by epistemic diligence in another area which is more important, or which has more consequences, so that we would not wish to describe it as self-deception, and perhaps would not even wish to describe it as negligence.

Another example: someone may be epistemically negligent through sheer laziness. He may be culpably deceived, yet not count as self-deceived. The difference between this kind of culpable error and self-deception, I think, is that a lazy person's motive is not linked to any particular belief or doxastic state. The lazy person in my example may indifferently count belief as good as ignorance, and false belief as good as true. Whereas, for a self-deceiver, the motive is connected to a particular belief (or the lack of that particular belief) and other beliefs would not be equally acceptable. That, anyway, is my intuition about how the word 'self-deception' is often used, and while there may be equally acceptable interpretations which differ from mine, it is the interpretation I am at work upon.

Negligence is only one of the strategies which a self-deceiver may adopt. It is so because it relies upon the situation being already set up so that just the belief required is generated, without any positive action from the self-deceiver, whose only role is to succumb to the charms of the situation.

A policy of mixed negligence and diligence would seem to offer more opportunities for self-deception. Negligence would be the response when the situation encourages the generation of a belief which suits the self-deceiver; otherwise diligence would be the response. The diligence might be diligent

criticism, to bring doubt upon an unwanted belief, or diligent generation of hypotheses, to introduce more options.

Introducing more options is a way of extending the self-deceiver's room for manouvre. For example, consider choosing a newspaper: if there is a wide range of newspapers to choose from, each representing a different slant and different ways of selecting what to report, then the self-deceiver can choose a newspaper which will tell him what he wants to hear. But there is a disadvantage as well: the self-deceiver may prefer there to be just one newspaper, provided it confirms his favoured belief (or lack of belief). For that would cut out the risk of being disillusioned.

The range of options might also be a range of exemplars of belief. Suppose person A has a belief; A might be said to have (wittingly or unwittingly) the recipe for that belief. An observer, person B, who wanted to believe what A believes, could use A as an exemplar: by aiming to put himself in the same position as A, and doing the same sort of things as A, person B may be able to successfully mimic A's recipe for belief, and so gain the belief.

Yet there may be no exemplars to follow, if the belief required is idiosyncratic. In that case the situation would not exist which allowed A to acquire his beliefs by negligence. Nor

could B gain beliefs by mimicry, copying a successful recipe for belief. Someone wishing to gain such an idiosyncratic belief would be obliged to invent the means to acquire the belief. Of course "the means" is the heart of the matter: it is the engine which generates the deception. Negligence is a contributory factor only insofar as it allows the deception to proceed. Self-deception by mimicry is a way of setting up the engine - the generator of the belief.

Negligence theories are also vulnerable to another kind of objection. Someone who is negligent may not bother to look for the evidence which would destroy his false belief; he may not think through the arguments which would make his belief untenable. But there are plenty of other people to do so on his behalf. They may present the evidence to him. They may go through the arguments with him. Sometimes he may be unable to avoid them. The false beliefs will then be liable to collapse under the onslaught of all the evidence and arguments which the self-deceiver did not bother to seek out. Even if he manages to avoid the confrontation, the very fact that he has avoided it indicates that he has been forced to alter his strategy. For instead of being passively ignorant of the evidence and the arguments, he is now actively ignoring them. Ignoring something is not at all the same as being ignorant of it: ignoring something involves being aware of something and

actively doing something to avoid being aware of it: we come back to the problems of Avoidance theories.

Negligence alone is not going to be sufficient to explain self-deception in more than a minority of cases. Negligence theories are adequate to explain only those cases of self-deception where the self-deceiver is protected from evidence and argument which would destroy the deception. When the self-deceiver has to do something to protect the deception, we are out of the realms of negligence. It takes more than an act of omission - it takes an act of commission - to sustain the self-deception in these circumstances. We are forced back to the claim that "the self-deceiver must know what he is up to", since that knowledge is needed to guide the strategy by which he "steers around the evidence" etc.

Negligence theories held the promise that we would be able to reject this claim, and thereby avoid the paradox of self-deception. The available negligence theories, in their current state of development, cannot deliver what they promised. To make them do so, we need to add two further elements.

The first thing we need to supplement negligence theories is a way of showing that "active" self-deception does not need to be guided by knowledge any more than "passive" self-deception

does. This element enables us to avoid one route to the paradox.

The second thing is to show how the self-deceiver is able to protect the deception against the coercive power of evidence and argument. Evidence and argument are supposed to be (sometimes, at least) "compelling", i.e. powerful enough to destroy one belief (such as the self-deceiver's favoured belief) and replace it with another (e.g. belief in something they make "obvious" or "evident"). This element is needed in order to deny the claim of the incredulous observer, that the self-deceiver "must" know something-or-other because it is obvious, and therefore the self-deceiver "must know that he is deceiving himself". In this way we can avoid another route to the paradox.

"Mere Pretence": Role Dissimulation

Role Dissimulation theories are Avoidance theories. They argue that the self-deceiver knows what he is doing, and that despite this knowledge, he is self-deceived. They claim that role playing is the technique used to deceive oneself. Role dissimulation is, as critics of the theories point out, "merely pretending". But the fact that someone is merely pretending does not rule out self-deception. By pretending, one can suspend disbelief (and suspend belief). By adopting a role, one can defer an unwelcome belief. Acting as if one does not have the belief, one need not use it to interpret data, nor need one use it to guide action; and instead of the belief evoking emotions, the emotions may be evoked by the role one is playing. So Role Dissimulation theories can explain how a self-deceiver "dissociates" or disconnects beliefs from actions, emotions, perceptions, and other beliefs.

However, someone who is self-deceived by dissimulating a role must also be constantly spoiling the self-deception. For he must be constantly refering back to the knowledge which guides the role, in order to steer the role and also, occasionally, to decide if the role is still worth sustaining. Also, if the role playing is periodic rather than continuous, he must decide

when to stop role-playing and when to start. This must disrupt the role-playing considerably.

If we suppose that dissimulation is a way of deceiving oneself, though, we may go on to ask why simulation is not also a way of deceiving oneself. For, as I argued in a preceding chapter, the difference between simulation and dissimulation is only a matter of degree: dissimulation is more restricted than simulation. As a result dissimulation forces the self-deceiver to continually start and stop the role playing as he moves between situations in which the restrictions do not apply, and situations in which they do. Simulation - because the role is not restricted - does not force the self-deceiver to continually start and stop.

The dissimulating self-deceiver may be able to prevent this stopping and starting, by avoiding situations in which the restrictions curtail the self-deception; but in order to avoid those situations, the self-deceiver needs to refer to the beliefs which the role was meant to disconnect and thereby disable. For otherwise he will not know which situations to avoid. However, if the self-deceiver lifts the restrictions upon the role playing - i.e. if he starts simulating rather than dissimulating - then he need not refer back to the unwelcome belief at all. For, despite the argument that he "must know what he denies, in order to steer round the

evidence", all that is needed is that he has some means of "steering around the evidence": he needs know-how rather than knowledge-that, and for this purpose a false belief may be as effective as a true one. Indeed, the history of enquiry suggests that a false belief is very often far more effective for "steering around the evidence".

Koestler [1968] provides some examples of (unwittingly) "steering around the evidence" by the use of false beliefs, false theories, false hypotheses. Koestler describes the long process of enquiry by which Kepler arrived at his Second Law of Planetary motion; Koestler then has this to say:

> by three incorrect steps and their even more incorrect defence, Kepler stumbled upon the correct law. It is perhaps the most amazing sleepwalking performance in the history of science - except for the manner in which he found his first law (p333)

Kepler unwittingly "steered around the evidence" for years; he did so by the use of false theories. Koestler again:

> At this point, the sleepwalker's intuition failed him, he seems to be overcome by dizziness, and clutches at the first prop he can find. ... and he falls back on the old

> quack remedy which he has just abjured, the conjuring up of an epicycle! (p334)
>
> To make the worthless hypothesis work, he temporarily repudiated his own, immortal Second Law - to no avail. Finally, a kind of snowblindness seemed to descend upon him: he held the solution in his hand without seeing it. (p335).

Koestler describes the activity of someone (Kepler) trying to gain the truth and being thwarted by the "worthless hypotheses" he uses. Imagine how effective a "worthless hypothesis" would be when used by someone who does not wish to gain the truth, indeed who cherishes a falsehood. Kepler's efforts also follow upon thousands of years when other, false, theories of planetary motion were adopted without any hint of being refuted by "the evidence". The evidence was "steered around" very successfully without the use of a true theory. Nobody had a true theory with which to "steer", and yet people did not notice, or did not recognise the implications of, "the evidence" which would have refuted their false theories.

It may or may not be possible for someone to be self-deceived by role dissimulation. Dissimulation could certainly be used to defer making use of a true belief that one has. Deferring the use of a true belief is a sort of self-deception - the sort

described by Dissociation theories as "believing something but not thinking it," or words to that effect. So Role Dissimulation theories can go some way towards supplying the details which are missing in Dissociation theories. Yet how much easier it would be to deceive oneself by simulation - which is much more belief-like and does not require the self-deceiver to "really know what he is doing". Role Simulation theories thereby seem to remove the paradoxes of self-deception at a stroke. Self-deception as characterised by Role Simulation theories is not so much mere pretence as sheer pretence: pretence which places false theories in the action-guiding and interpretive role which is performed by beliefs.

Before discussing Role Simulation theories further, I provide a chapter on Sartre's discussion of Bad Faith (Sartre [1975]). On my reading, Sartre characterises bad faith as a metastable process which "slides" between dissimulation ("cynicism") and simulation ("good faith"). However Sartre's characterisation can also be read as a description of role <u>dissimulation</u> only - the "slide" being between role-playing and momentarily dropping the role in order to refer to a guiding belief. My reading is intended to be the most "charitable" reading - attributing to Sartre the best theory I can read into his descriptions of bad faith. If Sartre's is a Role Dissimulation theory of self-deception, then it is vulnerable to the paradoxes of self-deception which haunt all Avoidance theories. For the self-

deceiver must have knowledge in order to deceive himself, and one would expect the knowledge to spoil the deception. Role Dissimulation theories are at their strongest in describing instances which can be plausibly described as deferring the use of one's beliefs. Where the self-deception does not seem like a deferring strategy, some other theory must be used. Role Simulation theories offer us this opportunity. Sartre, on my "charitable" reading describes how role dissimulation could be a stage in a process which leads towards role simulation. Someone who engages in role dissimulation is in the process of deceiving himself, and is on the way to being self-deceived; someone who engages in role simulation has achieved self-deception.

S a r t r e

I assume, along with other writers (e.g. Santoni [1978], Russell [1978], Morris [1980]), that Sartre's description of bad faith is a description of self-deception. Bad faith, he tells us, is a metastable state, "sliding" between good faith and cynicism; we could gloss this by saying that it is a movement between states rather than an achieved state. Nonetheless it can be long-lasting, and may even be a way of life for a great many people. A person in bad faith exploits the nature of consciousness in order to "flee from freedom".

Sartre makes a number of distinctions:

- between pre-reflective consciousness and reflective consciousness

    . Russell [1978] describes this as a difference between "immersing" and "detaching" consciousness; I discuss this further below

- between past, present and future aspects of the self within the temporal synthesis of consciousness

- between the first-person perspective ("being-for-itself" or "being-for-oneself") and a third-person perspective ("being-for-others")

- between different senses of the word 'I', namely:

  . my body (as in "I have blue eyes")

  . a summary of patterns of past behaviour (as in "I am a coward" - an attribution of character traits)

  . a chosen ideal self or "fundamental project" which gives structure, purpose and meaning to one's activities.

Morris [1980] remarks (correctly):

> Sartre has often been interpreted as a nihilist because of his claim that the human being begins as "nothing". It would be more accurate, I think, to see that Sartre is offering an activist version of the traditional empiricist's "blank tablet" view of man. (p36)

Sartre's distinctions allow us to describe a number of strategies for self-deception. The person in bad faith can trade off different aspects of selfhood in order to disavow

freedom and responsibility. For brevity, let us call one such person "B" (for Bad faith).

B can pretend that the present and the future are as decided as the past, by identifying wholly with his past self - thereby avoiding responsibility for present and future actions, since (according to the pretence) one's nature, or character, is fixed. The motto for this variety of faith might be, "you can't teach an old dog new tricks".

Another option is for B to ignore the "fundamental project" manifested by a pattern of actions in the past. B can own up to the actions, but treat them as isolated, trivial episodes, and refuse to see the pattern. B treats them as uncharacteristic deviations from the "ideal self" rather than as developments which cumulatively form his or her character and which, therefore, in practice constitute the "fundamental project". A motto for this variety of bad faith might be, "I wasn't myself - its so unlike me to do such a thing".

Garcin, a character in Sartre's play No Exit, exemplifies this strategy. He identifies himself with his future ideal self and denies the relevance of the third-personal perspective (the audience's perspective) which reveals his acts to be moving away from that ideal self (in Garcin's case, away from courage and towards cowardice).

Sartre's account is a Role theory of self-deception. We should note, though, that Sartre has a role theory of human nature: the role you adopt determines what you are. If you take on the role of a waiter, for example, then you are a waiter. Sartre describes a waiter who tries to be "nothing but a waiter" as being in bad faith. Not because he is pretending to be a waiter - he is a waiter - but by pretending that being a waiter is "in his nature", not something he could choose to alter. He thereby tries to evade his freedom to do (and be) something else instead, and shrugs off responsibility for his choices.

It might seem that in Sartre's account bad faith is "merely pretending", that the self-deceiver is not really deceived. But Sartre has a way of avoiding this objection. Firstly, we should question the "mereness" of "merely pretending". The person in bad faith is playing a role - but in Sartre's account we are all playing roles, with more or less conviction. Adopting a role is a way to constitute what one is. There is no "deeper" reality masked by the role: the role is the reality. This does not mean that by pretending to be Napoleon Bonaparte I can become Napoleon Bonaparte. It means that I am someone playing the role of Napoleon Bonaparte - there is no more genuine self hidden beneath or behind the role. In Sartre's example, by adopting the role of a waiter one becomes a waiter (of course, one must get the job in order to adopt the role - otherwise playing the role is indeed mere pretence).

Often we are not reflecting upon the role - we are "immersed" in it, as Russell puts it. But when we reflect upon the role, we become "detached" from it: instead of playing the role, we are thinking about it - adopting a different role. The person in bad faith "knows that believing always comes short of believing": consciousness of believing "spoils" belief. This is like the situation described by Palmer [1979]: one can make predictions about what someone else is going to do, but when one makes predictions about one's own future actions, the predictions cannot be distinguished from decisions.

I think that this is how Sartre regards reflection upon belief. By reflecting upon the belief, I detach myself from it, and I make it subject to my will. For if I choose to become something different from what I currently am then I become apt to believe things different from what I currently believe. Hence "belief always falls short of belief" (whilst we are reflecting upon it). So the person in bad faith decides that non-persuasion is constitutive of all convictions; "it [bad faith] accepts not believing what it believes". It plays a role of believing because, upon reflection, all beliefs are role-playing.

Playing the role, however, the person in bad faith is not all the time reflecting upon the performance. Instead of being "detached" he is "immersed": he becomes what he is pretending

to be - in this case, a believer. Now it suits the person in bad faith not to reflect upon the performance: reflection is dangerous because it contains the invitation to be something different. Sartre illustrates this with the example of someone feeling vertigo when looking down from a great height. The feeling arises because one knows one is free to jump from the height: the recognition of that freedom makes one dizzy. Another example, which I prefer, is the example of the gambler who has decided to give up gambling, and has resolutely told himself of this fact. Russell [1981] remarks,

> if he were to view with detachment his earlier pronouncements, he would thereby remove himself from that resolute person. So it is risky for him to reflect, with detachment, "You still could join the game." (p73)

Suppose he does so, and "succumbs": he goes towards the gaming tables; soon he is gambling. Now the dilemma is reversed: he could still walk away.

> Now he wants to avoid any commentary which would detach him from his sense of being swept up in the activities of the table (p73)

> that, I think, is part of the answer to why it would be important to him not to say [that he still could join the

> game]. One flees from explicitly taking the point of view of another, because one flees from being other than who one is choosing to be. (p73)

That is why bad faith is not "merely pretending": it is more than pretending because it is constituting oneself as a certain kind of person.

According to Sartre, bad faith is a precarious, "metastable" balancing act. The "sliding" between cynicism and good faith is a slide between reflective consciousness which decides that "all belief falls short of belief" and "immersing" consciousness when one is not reflecting upon the belief. But "good faith is still faith", Sartre remarks, suggesting that a person who is in good faith has not achieved Sartre's ideal of authenticity. Morris [1980] spells out the idea of authenticity implied in Sartre's Being And Nothingness:

> The person who does not deceive him- or herself is 1) one whose moments of reflection on his pre-reflective activities are accurate: he has learned to see his own activities as objectively as an outside observer would, when necessary, while still preserving a sense of his own goals; 2) one who not only knows what seperate acts he has done in the past, but can see what kind of pattern they form; 3) one who can make the correct connections between

> past acts and his ideal self - seeing where the acts obstruct, fall short of, or actually tend in the direction of the ideal; and 4) one who does not mistake that connection for a causal connection: he accepts responsibility for the fact that the present and/or future acts might follow a somewhat or even wholly different pattern from the pattern of his past acts, if he chooses a different ideal; he does not mistake that future self for a predetermined goal. (p44)

With the help of the other commentators quoted above I have completed a brief survey of Sartre's comments on bad faith. There are dangers in such exegesis. My rendition of bad faith may be a rather radical translation - even a mistranslation - of Sartre. But even if this is not Sartre's account of bad faith, it is certainly one account of self-deception, and one which most commentators seem agreed upon. So now let us ask what, if anything, is right or wrong with it.

Some elements of bad faith are familiar landmarks in the Anglo-saxon tradition. For example, "focussing" and "avoidance" are varieties of "immersing" consciousness which are often used to characterise self-deception. Immersing consciousness avoids reflection, avoids focussing attention upon the role the person is playing. Instead attention is focussed upon an object other

than the role. Reflective consciousness disarms commitment, and detaches the person from the role it focusses upon.

My view of self-deception differs from Sartre's account of bad faith in a number of ways. Firstly, in its moments of reflection bad faith (as described by Sartre) becomes cynicism: it decides to be "convinced when it is barely persuaded". In my account of self-deception, the self-deceiver need not become cynical when reflecting. For reflection, I suggest, consists of constructing an interpretation of the process of self-deception - an interpretation of an interpretation. This interpretation may be another misunderstanding of a misunderstanding. Detachment prevents one being immersed in the self-deception, but does not prevent one being deceived (in a detached, non-cynical way) about it.

Suppose this detachment takes the form of adopting a third-person perspective. Who is this third person? It may be someone objective (as Morris proposes); or it might be a deceiver, a self-deceiver, a tyrant, or someone just like oneself. Someone just like oneself will form just the same opinions, and so judge that the process is not self-deceptive.

On the other hand, suppose that this third person is someone very unlike oneself. The conclusion then may be that the process is self-deception: does the self-deceiver thereby

become cynical? He may do, he may not. He may come to the conclusion that, "I can understand why so-and-so would think that I am self-deceived, given his assumptions. But I do not share those assumptions, so I have no reason to suppose that I am a self-deceiver". The self-deceiver is not forced to accept the third-person perspective. To think that one was obliged to do so would be an instance of the attempt to flee from freedom and responsibility, and to make someone else responsible. Sartre himself condemns this attempt as a "flight from freedom" and "bad faith", so he is not entitled to claim that the failure to accept a third-personal opinion is bad faith.

If I am right then self-deception need not be metastable. It may not be a precarious balancing-act of directed attention, strategic immersion and detachment, and so on. In some circumstances a self-deceiver can share one characteristic of "authenticity", namely, "to see his own activities as objectively as an outside observer would, when necessary, while still preserving a sense of his own goals" (Morris, p46).

The self-deceiver I am talking about could also "accept responsibility for the fact that present and/or future acts might follow a somewhat or even wholly different pattern from the pattern of his past acts, if he chooses a different ideal; he does not mistake that future self for a predetermined goal". Probably he does not reflect accurately upon his pre-reflective

acts, or make the correct connections between past acts and his ideal self, or see what kind of pattern is formed by past acts - for after all, he is deceived about something. But the question of what is correct, what is accurate, etc, is up for grabs: it is not so obvious who is "authentic" and who is not "authentic".

I also have difficulty with some of Sartre's ontological apparatus. It may be just the difficulty of his terminology, style, love of paradox, etc. But (for example) I wonder if everyone has the "ideal future self" which allegedly gives structure to a life. Perhaps a lot of people have something which is a lot more ad hoc, such as that approach cited by Charles Reade:

> sow an act, and you reap a habit. Sow a habit, and you reap a character. Sow a character, and you reap a destiny.

The "ideal future self" has no role here. It does not exist even as an idea. A "self" is constructed as one goes along, perhaps in quite a haphazard way which is influenced by situations which arise independently of any conscious or unconscious plan of the person who will reap this destiny.

I intend to make use of Sartre's discussion as follows. The "slide" from "cynicism" to "good faith" can be mapped onto the change from role dissimulation to role simulation, or from "mere pretence" to "sheer" pretence. An instance of mere pretence is the construction of interpretations by "idle speculation", daydreaming, and so on: activities which can be performed as works of imagination without any commitment to believe the interpretations one constructs. That commitment is formed when one lifts the restrictions upon the pretence and starts to use it in the ways one would use a belief. This is done by adopting and sustaining a role. In cases of "sheer" pretence the role is not mere play-acting, it is the construction of a personal identity. It constructs a reality, not a sham or a facade masking and concealing a reality. For there is no other reality "behind" or "beneath" it. The interpretation it uses is the interpretation for that role, and since the role-player has no other, concealed role, it is the role-player's belief: his or her "genuine" interpretation. The role fixes the belief. It explains the role-player's aptitude to construct one kind of evidence rather than another, his aptitude to find some things "obvious".

The self-deceiver too can interpret freely, and then use an unrestricted role simulation to fix the interpretation and make it into a belief. Role simulation is the subject of the next chapter.

"Sheer Pretence" Role Simulation

> If the self-deceiver does not notice what he is doing then he is a case not of self-deception but of mistaken or biassed belief [...] If the self-deceiver is not really ignorant of what he is doing then he is a case not of self-deception but of a man who knows but simply pretends not to. (Mounce [1971], p66)

Mounce distinguishes self-deception from "simply pretending", and apparently thinks that one precludes the other. This cuts the ground from under Role theories before they have a chance to get started; but it does so without argument. So let us reinstate the opportunity to argue that self-deception is a variety of pretence, and argue from there.

A Role theorist might say of Mounce's observations that he has hit the nail on the head without noticing: self-deception is a variety of pretence, but not a simple one.

Role theories are hinted at in various places; one such place is Hamlyn [1971, (1)], which Mounce criticises. Another is the recipe for faith proposed by Pascal [1966] and others: act as if you have faith, and you will come to have faith. Pascal was

discussing ways of gaining faith when evidence is evenly balanced, not ways of deceiving oneself. However the recipe which is efficacious in that case may also be used as a means to self-deception when "the evidence" is not "evenly balanced".

Pretending that something is the case does not automatically lead one to believe that it is the case: but in some circumstances, with the right variety of pretence, it can. This is because all belief is a development of pretence: a pretence which is well-founded, we hope. I have not yet said which variety of pretence. So lets consider that now.

Pretence usually has limits imposed upon it. When NASA sent spacecraft on a "grand tour" of the solar system they used Newtonian physics to calculate the paths taken by the spacecraft. Yet they believed Newtonian physics to be false. They were using Newtonian physics as an instrument to generate predictions which were accurate enough for the practical purpose of plotting the course of the spacecraft. They were acting as if Newton's physics were true, in short they were pretending. Clearly this variety of pretence is not belief.

Theatregoers involved in a play often experience emotions rather like those which would be appropriate to someone who believes that the events portrayed on stage are real. They may also make deductions about the characters and about what is

going to happen next, in ways which would be appropriate if the events portrayed were real. The pretence they engage in, though belief-like in some respects, is unlike belief in others. The play takes place upon a stage, lets say: but the place portrayed upon the stage is probably not a stage, and the play is probably not about actors (there is no reason why it should not be, of course: it just so happens that most plays are not about actors and theatres). The pretence engaged in by the audience probably does not incorporate information about the lighted exit signs, the rows of seats, the footlights, etc: some of the available information is filtered out. Nor does it lead to belief-like behaviour: the audience does not leap up to disarm the actor on stage who simulates the murder of one of the other characters, for example.

Yet sometimes the conventional limitations of theatrical pretence are broken. Actors from the long-running serial Coronation Street have sometimes been accosted in the street and addressed as though they were the characters they portray: the viewer's pretence spilled over into "real life", beyond the boundaries conventionally imposed upon the viewing of television soap operas. When it "spills over", the pretence becomes more belief-like. Suppose that "overspill" of the pretence is extended. Suppose we progressively remove the limitations imposed upon the pretence - what then? I suggest that then what we arrive at is fully-fledged belief: which is

pretence upon which the pretender imposes no limit. Like the practical joke which "goes too far" and, ceasing to be a joke, "becomes serious", pretence can go beyond some conventional limit of pretence and become belief-like.

The pretender may not impose a limit. This does not mean that there is no limit: there may well be. But the limit is exactly that which we would expect a belief to have e.g. it is likely to fail if found to be inconsistent or turns out to be nonsense, provided that the person having the belief is committed to consistency and sense.

Someone who begins by making a wild guess at something may end up committed to its truth. At first he may "entertain it as a hypothesis", then start working out some of its implications; perhaps it starts to become more interesting: he discovers that it provides very neat explanations in an economical way, and clears up what had formerly seemed to be problems for any of the other available explanations. He brings it to bear upon tougher problems, and it works. He starts to think that it is true. At some point he comes to believe it; but believing it is a development of pretending it: it is what happens when a powerful pretence breaks through the limits imposed upon it. When he comes to believe it, does he stop doing the things he was doing when he was pretending? I suggest not. What makes something a "mere" pretence is that it is restricted: beyond

some limit it is not used. It collapses as we cease to suspend disbelief. In the case of the pretence which becomes belief, we never cease to suspend disbelief, but eventually there is no disbelief left to suspend. But the pretence could only become powerful and become a belief because we suspended disbelief (or because, being "gullible", we had no disbelief to suspend).

The self-deceiver plays upon the borders of belief and pretence: his pretences become beliefs. But this situation is not unique to self-deception: it is a way of getting beliefs even when there is no self-deception. Indeed, being unwilling to pretend is often taken to be a sign of self-deception ("he won't even entertain the hypothesis - he won't even consider the idea that it might be true"). Pretending is a way to move from one belief to another, and willingness to engage in some varieties of pretence is taken to be a sign of an "open mind".

So someone who pretends without imposing limits upon the pretence (a) acts like someone who believes (b) genuinely believes. The "man who believes but simply pretends not to" risks losing the belief he pretends not to have, and will lose it if the pretence is without imposed limits. For if there are no imposed limits then the pretence precludes him remembering that he is "only" pretending and that "really" he believes what he pretends not to believe. If the pretence is without limit then he cannot preserve the secret inner train of thought which

a liar (for example) must maintain. If the liar wishes to deceive others without deceiving himself, then he must preserve the secret,inner train of thought. If he does not preserve it, then he will fall victim to the deception, for he will forget what he "really" thinks, and be convinced by his behaviour just as other people are. The liar must be duplicitous, for singleness - integrity - will leave him sincere and as deceived as anyone else; he may even be deceived though nobody else is.

The deceiver of others dissimulates: the self-deceiver simulates. This does not mean that the simulation goes undetected by others. For the state of mind enjoyed (or endured) by the self-deceiver may be attainable in no other way than by such a simulation. The self-deceiver may be deceived through the lack of duplicity; but other people observing him are always in a "double" situation: there is the self-deceiver, and the observer, and they are two. The observer is not deceived by the integrity of the simulation, for he is not integral to the unity of the self-deceiver.

So we might propose as a rule that those who set out to deceive others are duplicitous, while those who deceive themselves "have integrity" - they are not divided enough to criticise themselves with an interpretation independent of the self-deception. This explains why self-criticism does not destroy the deception, but not how the deception is created.

Every belief, not just self-deception, is a simulation, a play, a pretence. And some forms of belief have pretence-like limitations. Someone may adopt a provisional belief, thinking: "I'll try this out, I'll gamble upon it working. If it works, well and good, I'll carry on with it. If it doesn't, I'll scrap the idea and think of something else".

As soon as we get into the game of provisional belief, our description of it becomes open to objections which show how unlike belief it has become. Someone who has such a lack of commitment to the "belief", one can argue, does not really believe at all. Yet he is not merely speculating. There is some commitment, a willingness to gamble something upon the truth of the "belief"; and if we follow this route, are not all beliefs, perhaps at some extreme limit, provisional? We gamble upon them, but there may be circumstances in which we would abandon even our most strongly-held beliefs. What is "provisional" and what is not, may be a matter of degree.

Role Simulation theories explain self-deception while escaping the problems and paradoxes of Role Dissimulation theories (and Avoidance theories in general). For they do not require that the self-deceiver "knows" what he is doing.

What is wrong with Role Simulation theories? My answer is, not much. However, what they rely upon is explanation by

assimilation: they suggest that self-deception can be understood by treating it as a variety of pretence. But does that make everything clear? I suggest not. For our understanding of pretence is not clear. "Simply pretending" is by no means as simple as Mounce's article might lead us to believe.

Also, we might assimilate in the reverse direction: instead of explaining belief as a variety of pretence, we might explain pretence as a variety of belief, or even as "suspension of disbelief". Better than either, we might consider some characteristics of both belief and pretence, and "contrast and compare", bringing them into a wider arena which includes belief, pretence, and other interesting human activities.

> the procedures by which we establish truth and falsity have, in the case of the self-deceiver, been tampered with by desire, so that what we have is a different game, parasitic upon the first, but differing from it at a number of points [...] since it is within the proper game of establishing truth or falsity that the terms knowledge and ignorance get their sense and since the self-deceiver does not play the game properly, one can say him, neither that he knows nor that he is ignorant. (Mounce [1971], p65)

It is often pleasant to renew acquaintance with an old friend, and here is one of our oldest friends, which I shall label "reason versus". In Plato it was reason versus the appetites; in Descartes it was reason versus those passions which led us to misuse our power to make judgements at will; here it is reason versus desire. Reason establishes the proper procedures for establishing truth and falsity, but desire tampers with them. All of which would be upset if reason were an expression of desire, or if there were such things as passionate reason or reasonable passions, or if reason were an appetite, one among many. Can desire "tamper" in something which it instigated in the first place? And if desire or passion or appetite did not instigate enquiry, is "reason" capable of doing so? I lump together desire, passion and appetite here, because I think that they play similar roles in all these accounts: they are a foil and a counterpoint for the ballet of reason, they are what need not be accountable or reasonable, for no-one is demanding that they should be. But desire has its logic just as reason has, and "reason", observed in actual instances of reasoning, can be thoroughly unreasonable, not because it has been "tampered" with by "external" forces, but because it never existed independently of them.

I take it that Mounce's "proper game of establishing truth and falsity" is the activity of enquiry. So Mounce's argument is that if someone is not engaged in enquiry then we cannot say of

them either that they know or that they are ignorant. This is a peculiar restriction upon how we are to talk. For it seems unexceptionable to ask if someone is ignorant or if they know, even if they are not "playing the proper game" of enquiry. We can ask whether the theatregoers know that what they are watching is a play and not an unscripted event in "real life": do they know, or are they taken in? They are not doing enquiry, I take it, yet they know that they are watching a play. They are not deceived by the pretence they are engaged in. So why is this instance all right but self-deception not all right? The question is rhetorical - they are both equally all right.

So what shall we say about self-deception - knowledge or ignorance?

> The self-deceiver will at one moment manifest what seems ot be knowledge and at the next what seems to be ignorance. From this we infer that he is moving from one category to the other ... the trouble is that on this interpretation the self-deceiver is not a self-deceiver in the normal sense at all. He is merely a man who is ignorant of what at a former time he knew. (p67)

Why all this mereness? Mounce goes on to argue that the semblance of knowledge is not real knowledge, and the semblance

of ignorance is not real ignorance. But I want to reply: the knowledge may be real knowledge; in cases of self-deception there is nothing "mere" about being "ignorant" of what one formerly knew: it is an achievement, like the old joke ("I wasn't born lazy, I had to work at it"). Some are born ignorant, some achieve ignorance, and some have ignorance thrust upon them: we shouldn't confuse the cases by inserting those little words 'mere' and 'simply' into our sentences.

But in any case I have reservations about the idea that someone who formerly knew can become ignorant. One may forget, but forgetting is not quite ignorance. One may mis-remember, but that kind of mistake - if it is a mistake -is not quite ignorance. Ignorance, once lost, is not so easily regained. But perhaps this is just a bit of legislation about how we should speak, and Mounce and I can agree to differ. Nonetheless I wanted to mark that difference between Mounce's use of words and mine, lest it misleads the reader elsewhere in this thesis.

The claim that believing is a variety of pretending is at odds with the claim that "belief at will is impossible". Williams [1973], Winters [1979], and Hampshire [1971] are among those who allege the impossibility of belief at will. Pretending is something one can do - more or less - at will, and believing is supposed not to be like that.

The word, 'pretend', certainly has an etymology capacious enough to hold both belief and deceit. Among the many meanings listed in the Shorter OED (Onions [1983]) are: to put forward, allege, claim, profess, to put oneself forward in some character, to feign to be or to do something, to feign in play, to make believe, to put forward as a reason or excuse, to use as a pretext, to allege; now esp. to allege or declare with intent to deceive (a leading current sense), from pre + tend or stretch.

Lets be clear about the status of this claim that "belief is a form of pretence". It is not a claim about how we use words, but a recommendation about how to use words, like the claim "whales are mammals" put forward at a time when whales were usually described as fish. The point is to draw attention to the continuities between pretence and belief, that what distinguishes belief from pretence is the limitations and conventions surrounding pretence.

Role Simulation theories are not-know theories of self-deception. They escape the paradoxes of self-deception because they do not require that the self-deceiver knows what he is up to nor do they require that the self-deceiver is "merely" pretending or "merely" ignorant. Unlike Negligence theories, they offer an account of what the self-deceiver does (namely, role-playing) rather than describing self-deception as a

failure to do something (namely, "proper" enquiry): they offer a positive account rather than a purely negative one. So far I have left this positive account at a rather abstract level, without giving examples of how role theories may be applied. However, the descriptions of Mr Negligent, Mr Mobile and Mr Radical, given in an earlier chapter, may give some indication of how role theories may be applied.

In the next few chapters I shall suggest how Radical Interpretation theories of self-deception can complement Role theories. Then I shall add an account of Marx's theory of ideology, which ties together the approaches of Role theories, Radical Interpretation theories, and my "Process theory" of self-deception. I shall then be in a position to describe the Process theory with a mimimal amount of subsidiary information. For the arguments in favour of the Process theory will have been presented in the chapters leading up to it.

A Different Metaphor: The Chemical Equilibrium

One of my motives for writing this thesis is dissatisfaction with some prevailing metaphors in epistemology. One might suppose that these metaphors are old, dead, and therefore inconsequential. I can show that this supposition is not correct by altering the metaphors and demonstrating the consequences. Metaphors may die but their influence lives on. For they give a structure to discourse which makes some questions and some answers obvious while "masking" others.

In this chapter I aim to replace a metaphor and indicate how this alteration alters the "obvious" questions and answers. This metaphor is the "scales of judgement". It describes judgement as an act of weighing. I shall replace it with the metaphor of the "chemical reaction". Both metaphors have a common theme. They are both concerned with the notion of equilibrium. I shall say something about each metaphor in turn.

1. The Scales Of Judgement

When the scales pans are empty, the scales of judgement are in equilibrium. When evidence is piled upon the scale pans, the equilibrium is disturbed and the scales tip to one side or the other. When the scales tip, a judgement is made, "coming down on one side of the argument".

The metaphor makes judgement into a mechanical matter. One may be selective about what evidence to put on the scales; one may avoid evidence, one may even fabricate evidence; but, once sufficient evidence is placed upon the scales its weight compels the judgement to take place. The evidence is "compelling". It exerts force upon the scales. The scales may waver before they tip. If the evidence weighs very little then there may be insufficient "weight" to tip them. If the evidence is evenly distributed between the two scale pans then the scales may stay in equilibrium and not tip. But whatever the outcome, once the evidence is on the scale pans, the operations of judgement are mechanical, like the operations of weighing.

a liar (for example) must maintain. If the liar wishes to deceive others without deceiving himself, then he must preserve the secret,inner train of thought. If he does not preserve it, then he will fall victim to the deception, for he will forget what he "really" thinks, and be convinced by his behaviour just as other people are. The liar must be duplicitous, for singleness - integrity - will leave him sincere and as deceived as anyone else; he may even be deceived though nobody else is.

The deceiver of others dissimulates: the self-deceiver simulates. This does not mean that the simulation goes undetected by others. For the state of mind enjoyed (or endured) by the self-deceiver may be attainable in no other way than by such a simulation. The self-deceiver may be deceived through the lack of duplicity; but other people observing him are always in a "double" situation: there is the self-deceiver, and the observer, and they are two. The observer is not deceived by the integrity of the simulation, for he is not integral to the unity of the self-deceiver.

So we might propose as a rule that those who set out to deceive others are duplicitous, while those who deceive themselves "have integrity" - they are not divided enough to criticise themselves with an interpretation independent of the self-deception. This explains why self-criticism does not destroy the deception, but not how the deception is created.

Every belief, not just self-deception, is a simulation, a play, a pretence. And some forms of belief have pretence-like limitations. Someone may adopt a provisional belief, thinking: "I'll try this out, I'll gamble upon it working. If it works, well and good, I'll carry on with it. If it doesn't, I'll scrap the idea and think of something else".

As soon as we get into the game of provisional belief, our description of it becomes open to objections which show how unlike belief it has become. Someone who has such a lack of commitment to the "belief", one can argue, does not really believe at all. Yet he is not merely speculating. There is some commitment, a willingness to gamble something upon the truth of the "belief"; and if we follow this route, are not all beliefs, perhaps at some extreme limit, provisional? We gamble upon them, but there may be circumstances in which we would abandon even our most strongly-held beliefs. What is "provisional" and what is not, may be a matter of degree.

Role Simulation theories explain self-deception while escaping the problems and paradoxes of Role Dissimulation theories (and Avoidance theories in general). For they do not require that the self-deceiver "knows" what he is doing.

What is wrong with Role Simulation theories? My answer is, not much. However, what they rely upon is explanation by

assimilation: they suggest that self-deception can be understood by treating it as a variety of pretence. But does that make everything clear? I suggest not. For our understanding of pretence is not clear. "Simply pretending" is by no means as simple as Mounce's article might lead us to believe.

Also, we might assimilate in the reverse direction: instead of explaining belief as a variety of pretence, we might explain pretence as a variety of belief, or even as "suspension of disbelief". Better than either, we might consider some characteristics of both belief and pretence, and "contrast and compare", bringing them into a wider arena which includes belief, pretence, and other interesting human activities.

> the procedures by which we establish truth and falsity have, in the case of the self-deceiver, been tampered with by desire, so that what we have is a different game, parasitic upon the first, but differing from it at a number of points [...] since it is within the proper game of establishing truth or falsity that the terms knowledge and ignorance get their sense and since the self-deceiver does not play the game properly, one can say him, neither that he knows nor that he is ignorant. (Mounce [1971], p65)

It is often pleasant to renew acquaintance with an old friend, and here is one of our oldest friends, which I shall label "reason versus". In Plato it was reason versus the appetites; in Descartes it was reason versus those passions which led us to misuse our power to make judgements at will; here it is reason versus desire. Reason establishes the proper procedures for establishing truth and falsity, but desire tampers with them. All of which would be upset if reason were an expression of desire, or if there were such things as passionate reason or reasonable passions, or if reason were an appetite, one among many. Can desire "tamper" in something which it instigated in the first place? And if desire or passion or appetite did not instigate enquiry, is "reason" capable of doing so? I lump together desire, passion and appetite here, because I think that they play similar roles in all these accounts: they are a foil and a counterpoint for the ballet of reason, they are what need not be accountable or reasonable, for no-one is demanding that they should be. But desire has its logic just as reason has, and "reason", observed in actual instances of reasoning, can be thoroughly unreasonable, not because it has been "tampered" with by "external" forces, but because it never existed independently of them.

I take it that Mounce's "proper game of establishing truth and falsity" is the activity of enquiry. So Mounce's argument is that if someone is not engaged in enquiry then we cannot say of

them either that they know or that they are ignorant. This is a peculiar restriction upon how we are to talk. For it seems unexceptionable to ask if someone is ignorant or if they know, even if they are not "playing the proper game" of enquiry. We can ask whether the theatregoers know that what they are watching is a play and not an unscripted event in "real life": do they know, or are they taken in? They are not doing enquiry, I take it, yet they know that they are watching a play. They are not deceived by the pretence they are engaged in. So why is this instance all right but self-deception not all right? The question is rhetorical - they are both equally all right.

So what shall we say about self-deception - knowledge or ignorance?

> The self-deceiver will at one moment manifest what seems ot be knowledge and at the next what seems to be ignorance. From this we infer that he is moving from one category to the other ... the trouble is that on this interpretation the self-deceiver is not a self-deceiver in the normal sense at all. He is merely a man who is ignorant of what at a former time he knew. (p67)

Why all this mereness? Mounce goes on to argue that the semblance of knowledge is not real knowledge, and the semblance

of ignorance is not real ignorance. But I want to reply: the knowledge may be real knowledge; in cases of self-deception there is nothing "mere" about being "ignorant" of what one formerly knew: it is an achievement, like the old joke ("I wasn't born lazy, I had to work at it"). Some are born ignorant, some achieve ignorance, and some have ignorance thrust upon them: we shouldn't confuse the cases by inserting those little words 'mere' and 'simply' into our sentences.

But in any case I have reservations about the idea that someone who formerly knew can become ignorant. One may forget, but forgetting is not quite ignorance. One may mis-remember, but that kind of mistake - if it is a mistake -is not quite ignorance. Ignorance, once lost, is not so easily regained. But perhaps this is just a bit of legislation about how we should speak, and Mounce and I can agree to differ. Nonetheless I wanted to mark that difference between Mounce's use of words and mine, lest it misleads the reader elsewhere in this thesis.

The claim that believing is a variety of pretending is at odds with the claim that "belief at will is impossible". Williams [1973], Winters [1979], and Hampshire [1971] are among those who allege the impossibility of belief at will. Pretending is something one can do - more or less - at will, and believing is supposed not to be like that.

The word, 'pretend', certainly has an etymology capacious enough to hold both belief and deceit. Among the many meanings listed in the Shorter OED (Onions [1983]) are: to put forward, allege, claim, profess, to put oneself forward in some character, to feign to be or to do something, to feign in play, to make believe, to put forward as a reason or excuse, to use as a pretext, to allege; now esp. to allege or declare with intent to deceive (a leading current sense), from pre + tend or stretch.

Lets be clear about the status of this claim that "belief is a form of pretence". It is not a claim about how we use words, but a recommendation about how to use words, like the claim "whales are mammals" put forward at a time when whales were usually described as fish. The point is to draw attention to the continuities between pretence and belief, that what distinguishes belief from pretence is the limitations and conventions surrounding pretence.

Role Simulation theories are not-know theories of self-deception. They escape the paradoxes of self-deception because they do not require that the self-deceiver knows what he is up to nor do they require that the self-deceiver is "merely" pretending or "merely" ignorant. Unlike Negligence theories, they offer an account of what the self-deceiver does (namely, role-playing) rather than describing self-deception as a

failure to do something (namely, "proper" enquiry): they offer a positive account rather than a purely negative one. So far I have left this positive account at a rather abstract level, without giving examples of how role theories may be applied. However, the descriptions of Mr Negligent, Mr Mobile and Mr Radical, given in an earlier chapter, may give some indication of how role theories may be applied.

In the next few chapters I shall suggest how Radical Interpretation theories of self-deception can complement Role theories. Then I shall add an account of Marx's theory of ideology, which ties together the approaches of Role theories, Radical Interpretation theories, and my "Process theory" of self-deception. I shall then be in a position to describe the Process theory with a mimimal amount of subsidiary information. For the arguments in favour of the Process theory will have been presented in the chapters leading up to it.

A Different Metaphor: The Chemical Equilibrium

One of my motives for writing this thesis is dissatisfaction with some prevailing metaphors in epistemology. One might suppose that these metaphors are old, dead, and therefore inconsequential. I can show that this supposition is not correct by altering the metaphors and demonstrating the consequences. Metaphors may die but their influence lives on. For they give a structure to discourse which makes some questions and some answers obvious while "masking" others.

In this chapter I aim to replace a metaphor and indicate how this alteration alters the "obvious" questions and answers. This metaphor is the "scales of judgement". It describes judgement as an act of weighing. I shall replace it with the metaphor of the "chemical reaction". Both metaphors have a common theme. They are both concerned with the notion of equilibrium. I shall say something about each metaphor in turn.

1. The Scales Of Judgement

When the scales pans are empty, the scales of judgement are in equilibrium. When evidence is piled upon the scale pans, the equilibrium is disturbed and the scales tip to one side or the other. When the scales tip, a judgement is made, "coming down on one side of the argument".

The metaphor makes judgement into a mechanical matter. One may be selective about what evidence to put on the scales; one may avoid evidence, one may even fabricate evidence; but, once sufficient evidence is placed upon the scales its weight compels the judgement to take place. The evidence is "compelling". It exerts force upon the scales. The scales may waver before they tip. If the evidence weighs very little then there may be insufficient "weight" to tip them. If the evidence is evenly distributed between the two scale pans then the scales may stay in equilibrium and not tip. But whatever the outcome, once the evidence is on the scale pans, the operations of judgement are mechanical, like the operations of weighing.

2. The Chemical Reaction

A standard notation for a chemical reaction looks like this:

$$AB + CD \rightleftharpoons AC + BD$$

A, B, C and D are chemical elements or combinations of elements which can in turn be combined in various ways. The symbol '$\rightleftharpoons$' indicates an equilibrium between two ways of combining A, B, C and D. So, for example, if we mixed together AB and CD then some of the mixture would react together to form different combinations AC and BD. There would be an equilibrium between the mixtures shown on either side of the symbol '$\rightleftharpoons$'.

This equilibrium can be changed by various means. For example: altering the temperature or pressure under which the reaction takes place; introducing other substances into the mixture; or removing some combinations from the reaction as they are formed. So, for instance, if we remove AC and BD as they are formed, then some of the remaining combinations AB and CD will react together in order to "restore the equilibrium", forming more of AC and BD. If we continue to remove AC and BD then eventually we shall remove everything: we will have tipped the equilibrium so as to produce AC and BD only. Instead of speaking of the weight of evidence (as we do when using the metaphor of the scales), we may, using the chemical

metaphor,speak of the stability or instability of evidence under varying conditions. Substituting one sort of equilibrium metaphor for another may seem a small change, but it can have a considerable effect upon the way we think about judgements.

Altering the equilibrium of a chemical reaction is somewhat like "tipping the scales", but there is a difference. Provided there is sufficient evidence and it is not evenly distributed between the scale pans, the scales will automatically tip down, making a judgement about any evidence presented. The tipping of a chemical equilibrium is more complicated: the outcome depends upon many factors (temperature, pressure, presence of other substances, removal of substances as they are formed, etc). So there is a question which the "scales of judgement" metaphor helps us to overlook while the "chemical reaction" metaphor makes it almost unavoidable, namely: if one has acquired "the evidence", what else has to happen to alter the equilibrium so that one "comes down on one side" of an argument?

Also, the metaphor alters our picture of evidence. In the "scales of judgement" metaphor, evidence is placed upon one scale pan or the other and the contents of one pan does not interfere with what the other pan contains. This is not so in the "chemical reaction" metaphor. AC and BD (the evidence shown on one side of the equilibrium notation) come into

existence because some of AB and CD (the evidence on the other side) have ceased to exist. For AC and BD are made out of AB and CD. AB and CD are transformed into AC and BD. What this means is that when there is more than one interpretation of data, one interpretation is constructed at the expense of another. We cannot "weigh" one interpretation against another upon the scales of judgement, for they are not independent of each other: in the "chemical reaction" metaphor, one interpretation (AC + BD) eats up another (AB + CD). What we can do is to observe the equilibrium between the two interpretations. It becomes important to answer the question mentioned above, namely: what has to happen to alter the equilibrium so that one "comes down on one side" of an argument? This is important because the answer tells us how one makes a judgement - which the metaphors are intended to (figuratively) describe.

I have partially answered the question in previous chapters. As long as one is engaged in "idle speculation" with no urgent need to take action, one can alter the equilibrium at will: the interpretation is not fixed, one can waver between one interpretation and another. However, adopting a role and making use of the interpretation fixes it, making it belief-like. This is like using up AC and BD, removing them from the chemical reaction so that more of AB and CD are transformed

into AC and BD: ultimately the equilibrium is tipped entirely to one side.

The "scales of judgement" metaphor makes this seem like an unwarranted interference in the operations of judgement. Doing something to alter the equilibrium is like putting one's finger upon a scale-pan while weighing the evidence: it introduces bias. This consequence of using the "scales of judgement" metaphor disappears when we use the "chemical reaction" metaphor. For the chemical reaction metaphor does not imply that the operations of judgement would occur automatically, like the operations of weighing-scales when weighty objects are put into the scale pans. It does not imply that there is any natural equilibrium between two different interpretations of the same data, irrespective of circumstances. The tipping of the equilibrium depends upon many things other than the properties of the evidence (the "weight" of the evidence). Making up one's mind is not merely a matter of being swayed by the "weight" of evidence. It is more like deciding what the evidence is to be, what is to count as evidence. This is not an unwarranted activity but one which inevitably enters into the process of making a judgement.

This leads to another change in the sort of questions we ask. For instead of asking what it is about evidence that compels belief, we may ask what it is about us which makes us apt to

interpret data one way rather than another. I suggest that this opens up a much more fruitful line of enquiry than the claim that belief is compelled by a property of evidence, namely "weight" - whatever that may be.

The chemical reaction metaphor also enables us to remove some questions which the scales of judgement metaphor invites us to ask. For example, how do we know which scale-pan to place evidence upon when we "weigh" it - in other words, how do we know to which side of the argument an item of evidence belongs? In the chemical reaction metaphor the question is answered automatically: because interpretations are not independent, items of evidence can only exist within their "side" of the equilibrium: they cannot float over to the other side of the equilibrium because moving to the other side of the equilibrium destroys them and converts them into something else.

I have sketched out how a change of metaphor can alter the questions we ask and the answers we give. I have also indicated how the "scales of judgement" metaphor, in particular, can (and does) influence the epistemological treatment of judgement, a process by which we arrive at beliefs. In doing so, I have set the scene for the following chapter, in which I discuss "Radical Interpretation" theories of self-deception. The change of metaphor shifts the emphasis from the question, "what is it about evidence that compels

belief?" to, "what is it about us that makes us apt to construct evidence in one way rather than another?".

I do not want to forget that the "chemical reaction" metaphor is only a metaphor. By this I mean that it is only one metaphor among others, not that it is "merely" a metaphor and therefore in some way inferior to a more "literal" description of the way we acquire beliefs. Although we can and often do try to give literal descriptions in place of metaphorical ones, a literal description does not replace a metaphor: it translates a metaphor. Our metaphors are never replaced (except by other metaphors). Literal descriptions preserve the metaphors which they translate, for the metaphors (however "dead" and "merely decorative" they may seem to be) remain active and effective, guiding discourse and determining which questions we ask and answer, which questions we notice and which questions we overlook (because the metaphor prevents the question arising or seems to answer it automatically).

The role of metaphors in the process of enquiry is an important one for radical interpretation theories. For example Nietzsche, that most radical of radical interpreters, argues that,

> truths are illusions of which one has forgotten that they <u>are</u> illusions; worn-out metaphors which have become

> powerless to affect the senses; coins which have their obverse effaced and now are no longer of account as coins but merely as metal. (Nietzsche [1964], "On Truth and Falsity in Their Ultramoral Sense", p80)

Nietzsche is not removing the distinction between truths and falsehoods here: he is not arguing that truths are falsehoods. Glossing what Nietzsche writes, one could say: truths and falsehoods are both illusions: but truths are reliable illusions, falsehoods are unreliable illusions. True (cf 'troth') is still distinguished from falsehood (cf 'fail'). Or one could say, with the physicist Niels Bohr,

> there are two kinds of truth, small truth and great truth. You can recognise a small truth because its opposite is a falsehood. The opposite of a great truth is another great truth. (Niels Bohr, quoted in Von Oech [1990], p119)

"Weaving": Radical Interpretation

> Discovery consists of looking at the same thing as everyone else and thinking something different. (Albert Szent-Gyorgyi, quoted in Von Oech [1990], p7)

The ability to "discover" by innovative thinking may also be used for the purposes of self-deception. People who "think something different" from everyone else are idiosyncratic, radical interpreters. They run the risk of being misunderstood, dubbed as self-deceivers, charlatans, or liars. If the interpretation is radical enough then they are likely to be told that they do not understand the meanings of words, that they are abusing language, that what they have to say is "merely figurative" and not to be understood literally. They are transgressors, acting against established norms (for otherwise they have not been very innovative). Furthermore, from the standpoint of the norms which they have transgressed, all these allegations may be true. If there are norms of discourse which fix what can be said or thought, then by transgressing them the radical interpreter speaks unintelligably or falsely.

Since (allegedly) evidence compels belief, and since evidence is constructed by a process of interpretation, there is a very simple and obvious way to change one's beliefs: change the interpretation. With this powerful ability one needs no elaborate strategies in order to be self-deceived. There is no need to avoid data, one needs no special acts of mentation to ignore evidence that one knows or believes. One needs only the normal processes by which beliefs are acquired. "The" evidence will not trouble the self-deceiver, since he does not perform the process of interpretation needed to construct it. Instead he constructs another interpretation which, though false, can be instrumentally effective - not effective for all purposes, but effective for his specific purposes. This interpretation can then be fixed by adopting a role. Used as a belief, it becomes more like a belief and can become exactly like a belief. Being exactly like a belief, it is a belief.

Understood in this way, self-deception is not paradoxical. For there is nothing paradoxical about fixing an instrumental understanding by using it as one would use a belief - which is the self-deceiver's "strategy". If anyone still objects to my calling this "belief", then I am prepared to give them the word 'belief'. Much good may it do them. For they will have divorced the word from the normal processes by which we acquire beliefs - the processes which I have described.

Avoidance theorists argue that to do what I have described is "merely pretending" to believe. Role theories provide an answer to the objection. The role one adopts fixes the interpretation one uses, and a fixed interpretation is a belief. We can stop fixing an interpretation by shedding the role.

In an earlier chapter I described three "strategies" of self-deception. These were the strategies of Mr Negligent, Mr Mobile, and Mr Radical, who fix their beliefs by sustaining, adopting, or inventing a role. These strategies are not unique to self-deception, however. It is not always self-deceptive to fix one's beliefs by one's role; the beliefs may be true, after all.

In this chapter I discuss what roles fix: interpretations.

Mr Negligent would assuredly be among those who argue that beliefs are generated by "the" evidence, that one cannot believe at will, and, therefore, that Mr Radical's efforts are "merely pretending to believe". For "the" evidence, i.e. the evidence which is already available, is constructed by "the" interpretation, i.e. the one which is already established and which it suits Mr Negligent to use.

Mr Radical's reply is that when Mr Negligent argues against radical interpretation he is arguing against himself. For Mr Negligent's own interpetation was, once upon a time, constructed by the process which he now castigates, namely radical interpretation. His appeal to "the" evidence is a way of giving preference to an interpretation solely because it is established, while making out that it is not an interpretation at all.

Self-deception proceeds by the normal strategies for acquiring beliefs, but it does so in a way which is not truth-regarding. The process of interpretation is a way of identifying, selecting, and ordering data. Different interpretations do this in different ways. So for example if A and B are two people using rival interpretations, then A ignores and disorders data which B identifies, while B ignores and disorders data which A identifies. One interpretation is used to organise data at the expense of another which it disorders.

To A, B will seem to be using a way of misunderstanding; and since A is not using B's interpretation, he will not understand by means of B. Likewise, to B, A will seem to misunderstand. B may even argue that A aims to misunderstand, i.e. that he is self-deceived.

My point is that self-deception is not merely a way of misunderstanding. It is also a (false, but instrumentally effective) way of understanding, with the aim of achieving some goal. Other people may regard the self-deceiver as aiming to misunderstand, but from his own point of view what he is doing is to understand in an effective way. The self-deceiver is not merely ignorant, nor merely mistaken, nor merely pretending, nor does he "really" (deep down) have true beliefs or knowledge which he is masking or avoiding. Self-deception is none of these things. It is false understanding, adopted because it is instrumentally effective for the self-deceiver's purposes.

Nietzsche makes the same point but reverses the stress. I argued that the self-deceiver's "misunderstanding" is also a kind of understanding. Nietzsche argues that understanding is also a way of misunderstanding. By doing so he calls into question the whole enterprise of truth-seeking, arguing that, "the will to truth is a will to error", "perhaps our truths are only our unrefutable errors", and so on.

Nietzsche makes some interesting remarks about this state of affairs. I shall summarise some of these remarks and then defend them in detail.

(a) The argument goes as follows. Every interpretation - every way of understanding - is also a way of misunderstanding.

To pursue a gregarious and social way of life, we need agreements, rules of discourse by which to understand one another. Therefore there are conventions of public language which govern what can be said - and also, what can be thought, and what can be counted as true. If we break these rules we run several risks. The greatest of these is the risk of madness: we may deprive ourselves of any stable means of understanding. Also, by contravening the rules we risk being identified as liars (people who abuse the conventions in a harmful way).

However, the rules make some things unsayable and unthinkable; and this too may be harmful. For every interpretation - every means of understanding - is also a way of misunderstanding. Not only does it hide something from us, it also causes us to forget that it is hidden by our own activities - for we are the interpreters, radical or conventional. When we follow the conventions in order to understand, we forget that we are also, simultaneously, misunderstanding.

Furthermore, the conventional rules could not have been created by rule-following. They must have been made by rule-making in the absence of rules, or by rule-breaking if there were other, earlier rules which have been replaced. Conventional interpretations originate in radical interpretations.

Nietzsche offers a history of how radical interpretations are conventionalised so that we "forget" their origins, forget that they begin - and continue - not as "reason" but as "faith". They are not justified in advance: we "leap" to them. Once an interpretation is in place, we can construct a justifying discourse. The interpretation provides the mechanism by which a justification can be created. The justification works because it presupposes the preliminary and still unjustified work done by radical interpretation: it is constructed by "reason" working upon the results of "faith".

If the conventions of ordinary language force us to misunderstand, then so much the worse for ordinary language: we shall have to use extraordinary language instead. If we are not to be caught within the misunderstandings created by a single interpretation then we must be able to weave between interpretations. One interpretation enables us to notice and understand what another interpretation makes indetectable or, if detectable, then unintelligable. Nietzsche commends the use of a "plural style" - which he describes as a combination of "scepticism and arrogance"; such a style enables us to construct interpretations ("arrogance") and to dismantle them again ("scepticism").

It is worth recalling that Sartre describes this combination as "bad faith" - "sliding" between "cynicism" and "good faith".

Most commentators have taken Sartre's account of bad faith to be a description of self-deception. Nietzsche's own position seems to be that self-deception inevitably accompanies truth-seeking, so that he is only recognising a situation which applies to all those who seek truth: truth-seeking and self-deception are inextricably bound together: "the will to truth is a will to deception". We interpret, so we must be engaged in the dual process which is both understanding and misunderstanding. We cannot help forgetting the status of our truths (metaphors) and that we are their artistic creators; but we can make this forgetting active. "Sliding" is a way of doing so, for Nietzsche's "sliding" is also a "weaving" between one interpretation and another. Nietzsche's argument leads to the conclusion that the desire to stabilise a single interpretation is characteristic of self-deception. To avoid this we can weave between interpretations, dismantling deceptions even while (unavoidably) building new ones.

This outlook alters the goals one has in doing enquiry. Instead of aiming for established, secure truths, one will aim for a continuing process which destabilises errors.

(b) Now I want to elaborate upon this summary argument. I shall do so by picking out several themes in turn, discussing them, and at the end putting them back together. The themes are:

(1) Nietzsche's description of "madness and faith" - and (in my terminology rather than Nietzsche's), order and disorder;

(2) Nietzsche's claim that convention fixes beliefs and causes us to forget that they arise from processes of interpretation

(3) thinking the "unthinkable"

(4) Nietzsche's description of the phases in the life cycle of a metaphor

(5) Nietzsche's "will to truth"

(6) Nietzsche's strategy

(7) Nietzsche's "Forgetting".

Theme number (8) puts themes (1) through (7) together again.

(1) Nietzsche's "Madness" and "Faith": Order and Disorder

We have to stabilise our interpretations or risk madness, Nietzsche argues.

> The greatest danger that always hovered over humanity and still hovers over it is the eruption of madness - which means the eruption of arbitrariness in feeling, seeing, and hearing, the enjoyment of the mind's lack of discipline, the joy in human unreason. Not truth and certainty are the opposite of the world of the madman, but the universality and the universal binding force of a faith; in sum, the nonarbitrary character of judgements (Nietzsche [1974], p130)

The search for understanding is a search for order, the construction of an ordering. From another point of view, it is a disordering. Understanding is achieved at the cost of another misunderstanding. Unable to stop interpreting - for to do so is madness - we seem bound to cling to our interpretations. What we can do, though, is to weave through several interpretations, allowing one to supply the lack in another.

Here is an illustration of the claim that an interpretation distributes order and disorder. Suppose that we are plotting a graph. We may be able to plot our data as a straight line graph providing that we are willing to use a logarithmic scale: we can exhibit the orderly nature of the line by making the scale less obviously orderly. Einstein's wonderful question ("if you could sit on a ray of light, what would you see?")

ordered things relative to the speed of light, taken as a constant, so that space and time are interpreted as "relative"; Newton's mechanics ordered things in a different way, so that space and time were absolutes and velocities relative - making the constant speed of light a "mere contingent fact". Different interpretations distribute order and disorder differently.

We only understand what is orderly. So if order (an ordering) is established by making something else disorderly, then interpretations not only enable us to understand but also enable us to misunderstand (or: disable us so that we cannot understand). A self-deceiver interprets; he thereby is able not to know what bystanders think he "must know", and to misunderstand what they think he "must understand". There are no special strategies of self-deception, only special aims and motives which the self-deceiver has for doing what we all do, namely interpreting.

"Look at things this way and you can understand how orderly they are". So if you look at things a different way (choose a different scale for the graph, choose different physical constants) you cannot see order - or only order of a different sort. Since we only ever see the order, we are liable to assert: that is how the world is - rather than: that is what we have made of it. Thereby asserting not only that there is

an orderly interpretation, but that beyond it there is an orderly world to justify the interpretation.

> he has distinctly convinced himself of the eternal rigidity, omnipresence, and infallibility of nature's laws: he has arrived at the conclusion that as far as we can penetrate the heights of the telescopic and the depths of the microscopic world, everything is quite secure, complete, infinite, determined, and continuous. Science will have to dig in these shafts eternally and successfully and all things found are sure to have to harmonise and not to contradict one another. How little does this resemble a product of fancy, for if it were it would necessarily betray somewhere its nature of appearance and unreality. ("On Truth and Falsity in their Ultramoral Sense", Nietzsche [1964], III.ii. 379)

Nietzsche, however, argues it <u>is</u> a product of fancy. Our truths are instrumental, we use them because they work, and we abandon them when they do not work.

> if each of us had for himself a different sensibility, if we ourselves were only able to perceive sometimes as a bird, sometimes as a worm, sometimes as a plant, or if one of us saw the same stimulus as red, another as blue, if a third person even perceived it as a tone, then nobody

> would talk of such an orderliness of nature, but would conceive of her only as an extremely subjective structure. (Nietzsche [1964], III.ii.379)

> All obedience to law which impresses us so forcibly in the orbits of the stars and in chemical processes coincides at the bottom with those qualities which we ourselves attach to those things, so that it is we who thereby make the imnpression upon ourselves. (Nietzsche [1964], III.ii.380)

> waking man per se is only clear about his being awake through the rigid and orderly woof of ideas, and it is for this very reason that he sometimes comes to believe he was dreaming when the woof of ideas has for a moment been torn by Art. (Nietzsche [1964], III.ii.381)

Stressing the role of interpretation does not commit one to the claim that all interpretations are equally good or equally bad. We do not have to like any one particular ordering, especially if we suffer the disorder it wreaks:

> the institutions of a society may well be at odds with its norms. To maintain these institutions may be fatally destructive. ... Theorists become revolutionaries only when their theories are able to articulate a deep dissatisfaction which the theorists did not invent. And

> at this point it is the refusal to destroy and recreate social institutions which is destructive of social life itself. The true nihilists in history were all kings: Charles I, Louis XVI, and Tsar Nicholas. The revolutionaries in their societies had to save social life from their rulers' destructive maintenance of the existing order. (MacIntyre [1967], p229 - 230)

In doing so, they articulated movements which disordered, disrupted, and remade not only the social institutions but the ways of understanding available within their societies.

(2) <u>According To Nietzsche, Convention Fixes Beliefs</u>

Nietzsche argues that interpretations are constructed by figuration. Some interpretations are fixed by convention. When the conventions become habitual we "forget" that they are invented. We regard them as being beyond our power to change ("belief at will is impossible") but what fixes them is the massive agreement of public institutions which coerce us into consenting, and the attempt by the supporters of those institutions to monopolise the process of understanding by insisting that it can proceed only by means of the instituted norms. Nietzsche's courage is to speak against those institutions, defending the task of thinking what those conventions would make unthinkable. The next section gives an

example of how an interpretation can make something "unthinkable".

(3) Thinking The Unthinkable

At one time it was illegal to belong to a trade union in Britain. Those who controlled the laws used their power to define trade unionists as criminals, by making the Combination Acts into law. Had they been able to control the moral vocabulary too then they could have made it true by definition that criminals should be prosecuted. Having power to enforce the law, they could (and did) prosecute trade unionists.

Their power to define vocabularies would have denied their opponents (and themselves) any language in which to utter - (or think) their opponents' case. The case would therefore have become undetectable to them. The domination of the vocabulary would have made some things unthinkable and unsayable - which is precisely what a self-deceiver might hope to achieve when he tries to "mask" or "conceal" something which he does not want to believe.

The Combination Acts are no longer on the statute books; and the struggle to dominate the moral vocabulary (or vocabularies) is still in progress. So we are able to think differently from those legislators. But perhaps other struggles have been

fought and won, or have not yet begun: if so then there are things which are unspeakable and unthinkable in our language. What are they? I cannot say or think what they are. To do so, I would need some other language (in which case, who will understand me?) or I would need to proceed through paradoxes and contradictions - in which case I could be shown to be misusing words, twisting them, and uttering falsehoods.

Nietzsche argues for the opening of struggles to think what cannot (according to the prevailing social norms) be thought, or uttered:

> Is language the adequate expression of all realities? (Nietzsche [1964], III.ii.373)

He argues that knowledge, truth and belief are constructed by interpretation. Interpretation constructs justifications. It is a leap into faith, a faith which resists opposition because it defines discourse, making some things unspeakable and unthinkable. Some understandings are constructed only by speaking against the conventional manner, which is the truthful manner (for the conventions define what can be counted as truth).

(3) Phases In The Life-Cycle Of An Interpretation

Nietzsche describes successive phases in the processes by which metaphors become "truths". The phases are as follows:

Phase 1: the "artistic" creation of a metaphor.

> That impulse towards the formation of metaphors, that fundamental impulse of man, which we cannot reason away for one moment - for thereby we should reason away man himself - is in truth not defeated nor even subdued by the fact that out of its evaporated products, the ideas, a regular and rigid new world has been built as a stronghold for it ... This impulse constantly confuses the rubrics and cells of the ideas, by putting up new figures of speech, metaphors, metonymies. (Nietzsche [1964], III.ii.379)

Phase 2: In order to satisfy the human need for a social and gregarious mode of existence, we agree to all use the same metaphors. The metaphors are thereby conventionalised: in phase 3 they will become "common currency".

> man both from necessity and boredom wants to exist socially and gregariously, he must needs make peace and at least endeavour to cause the greatest bellum omnium contra

> omnes to disappear from his world ... that which henceforth is to be "truth" is now fixed; that is to say, a uniform valid and binding designation of things is invented and the legislation of language also gives the first laws of truth: since here, for the first time, originates the contrast between truth and falsity. (Nietzsche [1964], III.ii.372)

Phase 3: the conventionalised metaphors are then naturalised. They become "truths":

> truths are illusions of which one has forgotten that they are illusions; worn-out metaphors which have become powerless to affect the senses; coins which have their obverse effaced and now are no longer of account as coins but merely as metal. (Nietzsche [1964], III.ii.377)

The common currency, the shared metaphors, have become debased. The conventional value of the metaphors (like the value of a coin) has been displaced by a seemingly natural value (like the value of metal).

Phase 4: once the convention of truthfulness is established, an activity develops which is parasitic upon the convention: lying.

> The liar uses the valid designations, the words, in order to make the unreal distinction appear as real ... He abuses the fixed conventions by convenient substitution or even inversion of terms. If he does this in a selfish and moreover harmful fashion, society will no longer trust him but will even exclude him. (Nietzsche [1964], III.ii.372)

Phase 5: the conventions become habitual. One forgets that they are conventions. This is a double forgetting:

- forgetting that the truths are metaphors

- forgetfulness of oneself as an "artistically creating subject".

> Only by forgetting that primitive world of metaphors, only by the congelation and coagulation of an original mass of similes pouring forth as a fiery liquid out of the primal faculty of human fancy ... in short only by the fact that man forgets himself as an artistically creating subject: only by this does he live with some safety, repose and confidence. (Nietzsche [1964], III.ii.378)

Habituation makes this forgetfulness possible:

> The very relation of a nerve-stimulus to the produced percept is in itself no necessary one; but if the same percept has been reproduced millions of times and has been the inheritance of many successive generations of man, and in the end appears each time to all mankind as the result of the same cause, then it attains finally for man the same importance as if that relation between the original nerve-stimulus and the percept produced were a close relation of causality: just as a dream eternally repeated, would be perceived and judged as though real. But the congelation and coagulation of a metaphor does not at all guarantee the necessity and exclusive justification of that metaphor. (Nietzsche [1964], III.ii.379)

Nietzsche adds:

> by this very unconsciousness, by this very forgetting, he arrives at a sense for truth. Out of the antithesis, "liar", whom nobody trusts, whom all exclude, man demonstrates to himself the venerableness, reliability, usefulness of truth. (Nietzsche [1964], III.ii.376)

He could have added: the "illusions" of metaphor become guilty by association with the liar.

What would happen if we could emerge from our double forgetfulness and, recognising the conventions for what they are, reassert ourselves as creative artists and our truths as mutable, chosen instruments?

> if he does not mean to content himself with truth in the shape of tautology, that is, with empty husks, he will always obtain illusions instead of truth. (Nietzsche [1964], III.ii.373)

Nietzsche appears to recommend that we should embrace these illusions by making our "forgetting" active: by choosing to forget rather than letting forgetfulness befall us. We can recognise that we are the artistic creators of the illusions. We can realise our freedom to create without doing harm and without being harmed.

The driving force of the will to truth is the impulse to create metaphors, a fundamental impulse of man.

The construction of truth, of self, of responsibility, of society, not only makes illusions - it makes these illusions come true. There are selves, there are truths, we do live socially and gregariously. The illusion lies in our forgetfulness of how they are created, and in the way they are

taken to be objective rather than being understood (or misunderstood) "only in their relation to man".

(5) Nietzsche's "Will To Truth"

The "will to truth" is ambiguous in Nietzsche's texts. It is a will to discover truths using the public conventions; but it is also a will to utter and think something within that language, in defiance of the conventions. Someone who defied the Combination Acts and its supporting moral vocabulary might argue that trade unionists are not criminals, that they ought not to be prosecuted, but he would have been arguing against the legal and moral vocabularies, and therefore against the "truth". He might succeed in altering the law and the moral vocabulary, and thereby succeed in making something true which had been made false by convention. Upholders of that convention could resist that "misunderstanding" and "misuse of words".

If we understand 'truth' in the narrower sense (truth is constructed within a vocabulary), then what is true depends upon who has the power to define the vocabulary. In the broader sense, truth is extra-linguistic, and the search for truth is an attempt to make something speakable and thinkable even if it has always been ruled unspeakable and unthinkable. Nietzsche suggests that our "truths" (in the narrower sense)

may be only our unrefutable errors (i.e. not truths in the broader sense). An established vocabulary is vulnerable to the eruption of another vocabulary able to overthrow it by making the formerly unspeakable into a conventional truth. The defeated may rise up and seize power - including the power to define.

Challenges to established truths have always proceeded by appeal to a "higher" truth, aiming to show the inadequacy of the established truths, and thereby to establish the higher truth - a more secure establishment. But this new establishment will, in turn, be guarded against the eruption of unspeakable truths. The aim to establish truths then seems inherently self-deceptive, since it is content to leave something unspeakable, and unthinkable. An interpretation which does not leave anything unthinkable would be disordered, chaotic, madness. A rule must rule something out, for otherwise it is not a rule. In any case we can never say that our interpretations are secure against the eruption of something it makes unthinkable, for we cannot find out anything about the unthinkable without convention-breaking.

If we cannot think the unthinkable, then it seems we have no alternative to our (thinkable) interpretation. But it is only unthinkable while we respect the conventions of the

interpretation. We can break them. We can construct interpretations "at will".

Some radical interpretations gain agreement and become the conventional, shared ways of understanding. Mr Radical's interpretation, for example, may become orthodox in the future. The formerly radical interpretation becomes the property of a new Mr Negligent. So it turns out that radical interpretation and conventional interpretation are two phases of a single process. This undermines Mr Negligent's claim that radical interpretation is mere pretence; but it also undermines Mr Radical, since it suggests that what he aims for is to become the future Mr Negligent: his criticism of Mr Negligent turns into criticism of his own goal.

(6) Nietzsche's Strategy

To defer forgetfulness, Nietzsche tells us, the "genuine philosopher" plays "the dangerous game", putting himself at risk. He weaves between madness and faith, and so, in a way, weaves them together. The game combines "scepticism and arrogance". Scepticism destroys faith, arrogance constructs it.

Mounce [1971] notes that the self-deceiver plays upon the edges of the "proper games" of knowledge and ignorance (so that self-

deception cannot properly be described as either knowledge or ignorance). Nietzsche claims that is where the "genuine philosopher" is to be found.

We could argue that Nietzsche is merely self-deceived. Yet he tells us what he is doing. A self-deceiver ought not to be able to do that. Nietzsche is "merely" pretending to be "merely" deceived. If we assume so quickly that pretending to be deceived is genuine self-deception, then we have accepted the claim of role theories, that pretending to believe can become belief, and Nietzsche's claim that the will to truth is a will to deception. For Nietzsche is re-enacting by "active forgetfulness" the processes by which truths and knowledge are constructed. In bad faith he simulates good faith, the will to truth.

Via "scepticism and arrogance" Nietzsche describes the construction of truth and knowledge (and falsehood, error and ignorance) as an activity of figuration and forgetfulness.

(7) Nietzsche's "Forgetting"

Nietzsche puts in question all the prominent expressions which have been used in descriptions of self-deception: 'belief', 'truth', 'knowledge', 'will', 'responsibility', 'self' - perhaps the only word which escapes his corrosive scepticism is

'deception'. However he does not commend unmitigated scepticism:

> How wonderful and new and yet how gruesome and ironic I find my position vis-a-vis the whole of existence in the light of my insight! ... I suddenly woke up in the midst of this dream, but only to the consciousness that I am dreaming and that I must go on dreaming lest I perish - ... Among all these dreamers, I, too, who 'know', am dancing my dance. (Nietzsche [1974], p116)

We cannot help forgetting. What we can do is to make that forgetting active:

> To close the doors and windows of consciousness for a time; to remain undisturbed by the noise and struggle of our underworld of utility organs working with and against one another; a little quietness, a little tabula rasa of the consciousness, to make room for new things, above all for the nobler functions and functionaries, for regulation, foresight, premeditation (for our organism is oligarchically directed) - that is the purpose of active forgetfulness, which is like a doorkeeper, a preserver of psychic order, repose, and etiquette; so that it will be immediately obvious how there could be no happiness, no

> cheerfulness, no hope, no pride, no present, without forgetfulness. (Nietzsche [1969], p57-58)

To believe our truths, we must "forget" that they are the artifacts left by our own activity, interpretation. The paradoxes of self-deception are transformed into paradoxes of knowledge and belief in general: "the will to truth is a will to deception". But the way in which the paradoxes are introduced also dissolves them by drawing attention to the role of interpretation: using one interpretation to achieve order disrupts another.

When we find out that some belief is false, we realise that the belief was "fabricated": that we made it up (or somebody, a deceiver, made it up for us). But the beliefs we hold true are fabricated by exactly the same techniques. Otherwise we would have a very convenient way of distinguishing truths from falsehoods. Attempts to define a "method" for justifying beliefs have exactly that aim: to find the techniques by which to construct only truths, never falsehoods. I suggest that such a method has not been found, and if it were found it could only be justified by results. So if we can distinguish truths from falsehoods, we must do so by some means other than their method of construction.

(8) Bringing The Strands Together

Radical Interpretation theories give a non-paradoxical account of self-deception by committing a sort of philosophical heresy: they blur a sharp distinction. However, once we have blurred the distinction, we can sharpen it again, in a different (and, to my mind, better) way. The distinction is between interpretations - products of an arbitrary, free activity, performed at will - and belief - a doxastic attitude which is allegedly fixed by evidence and not acquired at will. Radical interpretation theories argue that beliefs are interpretations. It is possible to arrive at beliefs by radical interpretation because all beliefs are acquired by interpretation.

The lack of a stable interpretation, Nietzsche tells us, is madness. We cannot avoid madness by reasoning, for until we have a stable interpretation there is nothing to which reasoning can appeal. Even the law of non-contradiction, which assigns the value, 'false' to all sentences of the form, 'P and not-P', requires that 'P' identifies the same item in both of its occurrences: this is impossible without a stable interpretation. So, for the sake of gaining stability, we need to make a leap of faith, and interpret.

However, a stabilised interpretation has its dangers too. Using an interpretation does not enable us to detect the

disorder it creates, only the order. Sticking to one interpretation is dangerous because it does not enable us to know what we are missing. Self-deception exploits this. To mitigate the way in which interpretations disable us, we must challenge the stable interpretation. The means to do so is the "impulse to metaphor". We can construct another interpretation by the same means which enabled us to construct the first.

If we destabilise our interpretations, we risk madness. We also risk the penalties attached to convention-breaking: we may be dubbed liars or told that we are merely misunderstanding the conventions of the public language. Since these conventions govern the use of the word 'true', we may be accused of "falling away from the proper procedures for establishing truth and falsity", i.e., accused of negligent self-deception.

However the radical interpreter is not merely mistaken (i.e. using the conventions ineptly) nor is he merely lying (abusing the conventions): he is challenging the conventions. This is where Bohr's distinction between great truths and small truths comes in: the radical interpretation may be another "great truth" or it may be a falsehood. The aim of radical interpretation may be to achieve another understanding, or (like a self-deceiver), to achieve a misunderstanding of the conventional interpretation. We can test which is the case in

any instance by testing the interpreter's willingness to "weave" between interpretations. If he is unwilling to contemplate both interpretations then he is avoiding understanding. However, willingness to "weave" may indicate self-deception by "inconstancy": the practice labelled "using double standards" may be self-deception by inconstancy (or it may be a way of deceiving other people only), for the two ways of understanding are also two ways of misunderstanding.

Let us see how the arguments of this chapter fit into the overall theme of this thesis, namely giving an account of self-deception.

Self-deception seems to combine two epistemic states which are inconsistent with each other - such as knowledge and ignorance, or true belief and error, depending upon how one characterises self-deception.

This can lead us to be sceptical about whether self-deception ever occurs. No-Such-Thing theorists argue that every attempt to provide a non-paradoxical account of self-deception collapses: either it collapses into a description of mere mistake (and so it is not a description of self-deception) or it collapses into a description of mere pretence (and so, again, it is not a description of self-deception).

I argue that there are no special techniques of self-deception. For example, self-deception is not achieved by a schism which splits the self-deceiver into two or more parts.

Self-deception exploits some characteristics of the usual ways in which we arrive at beliefs. It is not some unusual or abnormal way of arriving at beliefs; but nor is it wholly explained by assimilating it to various mental operations ("focussing", "ignoring", etc) with which we are all familiar. For although these operations may be familiar, assimilating self-deception to them has the effect of making them seem as strange and paradoxical as self-deception. This may be no bad thing - perhaps we ought to be puzzled and astonished when we pay attention to these familiar operations - but it does not explain self-deception. If we think that it does, then our critical faculties have been successfully lulled to sleep. In that case then we have been offered - and have accepted - a "magic button". It is as if self-deception were explained by saying that it is achieved by pressing a button (labelled, "focussing", for example). We are left waiting for a description of the processes which take place when the button is pressed.

Dissociation theories appeal to such magic buttons. Negligence theories argue that self-deception is the failure to press such a magic button - a button labelled, "the norms of enquiry",

"the proper procedures for seeking truth", and so on. There is no agreement as to what these norms and procedures are, so we can hardly explain self-deception by gesturing towards them.

We should pay attention to the processes by which the magic buttons work. This could be interpreted as a transgression - namely, attempting within a philosophical thesis to do "armchair psychology". However, I do not think that philosophers are entitled to expect psychology to answer philosophical questions - and we could not yet offer a philosophical account showing that self-deception can be characterised in a non-paradoxical way. So we were still at the mercy of No-Such-Thing theories of self-deception. They too were still vulnerable to a non-paradoxical account of self-deception.

My aim, therefore, was to construct a philosophical (not psychological) account of self-deception, with the promise that "the proof of the pudding is in the eating"; i.e. once we constructed such an account we would be in a position to discuss whether or not it is philosophical - not before.

Radical interpretation (RI) theories allow us to go one step beyond the "magic buttons" which have blocked the explanation of self-deception. RI theorists argue that beliefs, both true beliefs and false, are constructed by a process of

interpretation. This process is often described with the qualifying word 'mere' - mere interpretation, mere supposition, mere pretence - and contrasted with genuine belief which is not "mere", which allegedly is not achievable at will, and which is compelled by the "weight" of evidence and the "force" of arguments. My response was to identify the magic buttons on offer. "Will" is the magic button which allows us to produce interpretations "at will"; evidence is the magic button which allegedly produces belief. Interpretation is notoriously "wilful", free, arbitrary, etc, yet, I have argued, evidence is constructed by interpretation. So we need not suppose that evidence is coercive.

Attempting to construct a philosophical account of self-deception is a way to wake ourselves up from "forgetfulness". For, to give a non-paradoxical account, we have to acknowledge the activities of radical interpreters. Otherwise we must be amazed when evidence, the allegedly irresistable force, encounters self-deception, the immovable object.

Interpretations construct orderings and disorderings of data. Self-deception can seem puzzling if we consider it to be only a construction of disorder, a way of not understanding. To make it less puzzling we need to consider what sort of order it constructs. This is a risk. We risk being "converted" to the self-deceiver's interpretation; and the explanation we produce

may seem to be a defence of self-deception, since by aiming to make self-deception intelligable we risk making it acceptable.

If we are willing to weave between interpretations, we may apear to think that all interpretations are equally good or equally bad. Yet "weaving" between interpretations need not prevent us having preferences. It may enable us to detect areas of interpretation which have been "blocked off", and to detect structures of pain, craving and clinging which can explain the "distortions" and "diversions" in the self-deceiver's thinking. Our own prefered interpretations may also be subject to the same influences (pain, craving and clinging).

A self-deceiver can "hide things" from himself or herself because every interpretation hides things. We cannot say what things our own interpretations hide from us except by "leaping" or "weaving" into another interpretation. The sorts of things that are hidden by self-deception are "unpalatable facts" such as one's faults, mortality, aims - and those of other people too. To make life tolerable, to avoid madness, to preserve a rewarding interpretation, a self-deceiver desires to leave some things hidden. Self-deception aims to stabilise an interpretation, to make it permanently immune to change. Truth-seeking continually disturbs this situation; the self-deceiver's "discovery of truth" is a continual pacification.

We can destabilise interpretations. For example, the "chemical reaction" metaphor of the previous chapter invites questions and answers different to those offered by the "scales of judgement" metaphor.

The instability of our moral discourse is another example of interpretations being disrupted by other, competing interpretations. If everyone had agreed upon a moral vocabulary then no-one would have questioned the status of moral judgements as descriptions, no-one would have argued that they are ("merely") prescriptions. The little word 'merely' is used to seperate areas of discourse (for example "merely pretending" is used to invite the addition, "and therefore not really believing").

Nietzsche argues that the ambiguous status of moral descriptions / prescriptions is matched by the ambiguous status of truths in general (hence he writes of "Truth and Falsity In Their Ultramoral Sense"). Truth and falsity are "ultramoral" because they are constructed, prescribed, by rules about what it is permissable to say. The power to prescribe what may be said and thought is used to limit and constrain what is said and thought. Someone who exercises this power in an idiosyncratic way risks being labelled a self-deceiver, a charlatan, a liar, or being dismissed as quite mad; and the risk is also that the label may be true. But (this is

Nietzsche's anti-democratic impulse) the situation is no better when everyone agrees: the descriptions are still prescriptions.

The radical interpreter aims to make others understand something - and before that, to make himself or herself understand something - despite, and in contravention of, the currently prevailing norms of discourse. Figuration ("metaphor") is a means to do this without explicit paradox and self-contradiction.

The arguments presented here may not be intuitively convincing. However I am not in search of intuitive convictions. My aim is to construct a theory which works. If this aim is achieved, I do not much care whether the theory is intuitively obvious or counter-intuitive. Indeed, if our intuitions generate paradoxes, the theory had better be counter-intuitive. Part of my argument is that we construct "obviousness". If we start to use a counter-intuitive theory, then our intuitions will start to come into line with our theory. The question is whether or not this theory will do the work we want such a theory to do.

How shall we test our theory about radical interpretation? One way is to ask if it removes the paradoxes of self-deception. I

have argued that it does, and in later chapters I will test the theory against alleged examples of self-deception.

The radical interpretation theory fails if we can find a theory which does not originate in a process of radical interpretation. Since we cannot in practice work through every theory, determining its origins, the radical interpretation theory must remain a working hypothesis. In this thesis I can only give examples of how the hypothesis works. Radical interpretation theories remain vulnerable to a telling counter-example, if one exists.

In the next chapter I give an example of the history of some metaphors. I claim that this history exemplifies the conventionalisation and naturalisation of metaphor which Nietzsche outlines. These metaphors are used to construct models of memory and understanding which contribute to the argument that the self-deceiver "must know", that evidence is not a product of interpretation, and so on. Paying attention to the metaphors gives us a key to unlock the progress of epistemology, or at least a way of picking the lock.

The Mind's Waxy Essence

There are two metaphors which are used repeatedly over thousands of years of our history, and are so common that to remove them from the record would make huge chunks of that history unintelligable.

These two metaphors do not seem to be compatible with each other, as I shall explain below. By reconstructing a way of linking them together which makes them compatible, we can gain an insight into the way that "belief at will" - including self-deception - happens. We can also illustrate Nietzsche's description of the phases in the history of a metaphor.

The first metaphor is the "scales of judgement". This metaphor is used to understand judgement as a way of weighing evidence: evidence is placed upon the scale pans and the scales tips down on the side of the pan which bears the weightiest evidence.

The second metaphor is "the mind's waxy essence". This metaphor is used to suggest that the mind is malleable and retentive, like wax. It retains impressions of the things which strike it. The impressions are made by something weighty

(such as evidence) or something forceful (such as arguments). In particular, the wax retains "sense-impressions".

Here is a summarised history of the "wax" metaphor.

Cicero credits the metaphor to Simonides of Cheos (circa 556-468 BC), the "inventor of the art of memory":

> he inferred that persons desiring to train this faculty (of memory) must select places and form images of the things they wish to remember and store those images in those places ... and we shall employ the places and the images respectively as a wax writing-tablet and the letters written on it. (Cicero [1948], De Oratore, II, lxxxvi, 351-4)

It seems that Simonides intended to provide a model for training memory. He aimed to model mnemotechniques upon the best available technology for recording, which in his time was the technology of writing upon wax tablets.

The metaphor was put forward, elaborated and criticised, by Plato. Here is what Plato says:

> I would have you imagine, then, that there exists in the mind of man a block of wax, which is of different sizes in

different men; harder, moister, and having more or less purity in one than another ... let us say that this tablet is a gift of Memory, the mother of the Muses; and that when we wish to remember anything which we have seen, or heard, or thought in our own minds, we hold the wax to the perceptions and thoughts, and in that material we receive the impression from them as from the seal of a ring; and that we remember and know what is imprinted as long as the image lasts; but when the image is effaced or cannot be taken, then we forget and do not know ...

And the origin of truth and error is as follows: When the wax in the soul of anyone is deep and abundant, and smooth and perfectly tempered, then the impressions which pass through the senses and sink into the heart of the soul, as Homer says in a parable, menaing to indicate the likeness of the soul to wax - these, I say, being pure and clear and having a sufficient depth of wax, are also lasting; and minds, such as these, easily learn and easily retain, and are not liable to confusion, but have true thoughts, for they have plenty of room, and, having clear impressions of things, as we term them, quickly distribute them to their proper places on the block. Do you agree? [my emphasis].

> ... but when the heart of anyone is shaggy - a quality which the all-wise poet commends, or muddy and of impure wax, or very soft, or very hard, then there is a corresponding defect in the mind - the soft are good at learning, but apt to forget; and the hard are the reverse; the shaggy and rugged and gritty, or those who have an admixture of earth or dung in their composition, have the impressions indistinct, as also the hard, for there is no depth in them; and the soft, too, are indistinct, for their impressions are easily confused and effaced. Yet greater is the indistinctness when they are all jostled together in a little soul which has no room. These are the natures which have false opinion; for when they see or hear or think of anything, they are slow in assigning the right objects to the right impressions - in their stupidity they are apt to confuse them and are apt to see and hear and think amiss - and such men are said to be deceived in their knowledge of objects, and ignorant. (Plato [1949], 191 - 195).

Accepted by Aristotle, the metaphor becomes the standard, conventional model for memory. It is no longer put in question. It becomes established in a powerful theory about the way memory functions:

> When a stimulus occurs it imprints as it were a mould of the sense-affection exactly as a seal-ring acts in stamping ... memory does not occur in those who are in a rapid state of transition ... it is as if the stimulus, like the seal, were stamped on running water ... in others their worn-out condition ... and the hardness of the receptive structure, prevent the sense-impression from leaving an impression. (Aristotle, De Memoria, 450a - b, in Sorabji [1972])

By the mid-eighteenth century the metaphor has become so habitual and so "natural" that Hume can talk about "impressions" without noticing any need to tell us that the impressions are in (not wax but) the mind. It seems to be forgotten that the metaphor is a metaphor. It has become conventionalised, naturalised, literalised, so that Hume's use of the word conforms, perhaps, to the primary, literal meaning of the word, and its original meaning has become secondary (but also literal). The word is now used as a non-figurative expression.

Notice that this history matches what I called "phases in the life of a metaphor", with a high degree of precision. The metaphor ("the mind receives impressions") has become a conventional, obvious truth - true by definition, because the

use of words like 'impression' have come to organise the ways in which we talk about the mind.

Going back a few years, we find that Locke uses the metaphor to describe the mind as a "tabula rasa", a blank tablet of wax which receives sense-impressions. Locke [1964] links the metaphor of the wax to the metaphor of the scales of judgement.

Putting these metaphors together provides a way of describing the operations of cognition: the wax receives and retains impressions - like perception and memory - while the scales weighs up evidence - like judging something to be true or false. This combination of metaphors seems to leave little room for volition in the processes by which we gain beliefs. However, a third element is needed to harness the two metaphors together. For suppose that I use my senses to acquire evidence: I position myself so that something makes a sense-impression upon the malleable, retentive wax of my mind. But impressions do not seem to be the sort of things which can then be weighed upon the scales of judgement: impressions are better described as absences of weight, the shape of where the wax used to be. So we need to explain how the sense-impression is linked to the scales of judgement.

Without this missing link the metaphors fail to mesh together, and we can discern the symptoms of this failure in the

difficulties Locke has in his dispute with Leibniz. Locke denies that there are "innate ideas" in the mind, whereas Leibniz claims that there are. Locke argues that ideas are "abstracted" from sense impressions, and that therefore we have no need to postulate the existence of innate ideas. But what is the nature of this strange process, "abstraction"? Impressions are particular: how can we "abstract" from them an idea which is not particular but general? Supposing we can "abstract" ideas, how can we put them together to form something which can be "weighed", something which is capable of being true or false? Abstracted, generalised impressions (i.e. ideas), whether taken singly or put together in clumps, do not seem to fit the requirement: a bundle of ideas is just a bundle of ideas, not a truth-bearer.

Locke's problems can be traced back to the "missing link" between the wax and the scales. Descartes too uses the wax and scales metaphors, but he links them together via a third metaphor, the metaphor of the template.

Descartes argues that the understanding can act both as wax and as a seal - i.e. the template which forms the impressions in the wax:

> in all these processes, the cognitive power is sometimes passive, sometimes active; it plays the part now of the

> seal, now of the wax; here, however, these expressions must be taken as purely analogical, for there is nothing quite like this among corporeal objects. The cognitive power is always one and the same; if it applies itself, along with the imagination, to the common sensibility, it is said to see, feel, etc; if it applies itself to the imagination alone, as far as that is already provided with various images, it is said to remember; if it does this in order to form new images, it is said to imagine or conceive; if, finally, it acts by itself, it is said to understand. (Descartes [1970], p169)

The understanding can be formed into a template which fits many impressions - perhaps like a sculpture which can make many different impressions, and can be fitted into the many different impressions. Such a template would organise the many impressions by showing how they can all fit different parts of the template.

Weighing-scales cannot tip themselves, but we can incline them by putting weight upon them. We can construct templates to fit the impressions: the templates can then incline the scales of judgement one way or the other, by their weight. For this reason Descartes can argue that the "inclination of the will" can be both caused by the weight of evidence and none the less free:

> there is no need for me to be impelled both ways in order to be free; on the contrary, the more I am inclined one way - either because I have clearly understood it under the aspect of truth and goodness, or because God has disposed my inmost consciousness - the more freely do I choose that way
>
> however much I may be drawn one way by probable conjectures, the mere knowledge that they are only conjectures and not certain and indubitable reasons, is enough to incline my assent the other way (Fourth Meditation, Descartes [1968]).

Descartes' method for enquiry arises from this organised construction of the metaphors. He offers: (a) a mathematical method for the construction of templates and (b) a way to avoid errors: we do so by not constructing templates ("ideas") which are not "clear and distinct".

"Clear and distinct" has been taken to be the symptom of yet another metaphor, the "optical metaphor", and so it is. Descartes often writes of "the light of reason", for example:

> upon a great illumination of the intellect there follows a great illumination of the will (Descartes [1968])

But his understanding of light converts it into a tactile phenomenon, so that the optical metaphor is guided by the tactile characteristics of the primary operation of the understanding - the interplay of the wax and the template. There is no doubt that Descartes does use the optical metaphor, writing of "seeing by the light of reason" etc. But consider what he writes on the subject of vision:

> let us not deny anyone else's view of colour, but let us abstract from all aspects except shape, and conceive the difference between white, red, blue, etc., as being like the difference between shapes such as these: [and Descartes offers some drawings of shapes] (Descartes [1970], p167)

On "the external senses" he writes that,

> their having sensation is properly something passive, just like the shape (figuram) that wax gets from a seal. You must not think that this expression is just an analogy; the external shape of the sentient organ must be regarded as really changed by the object (Descartes [1970], p166)

and:

> So also for the other senses. The first opaque part of the eye receives an image (figuram) (p167).

There we have it: images are shapes. The metaphor of the impression in wax is Descartes' explanation of the optical metaphor.

One might argue that all this happened long ago, that since then we have outgrown such metaphors and learned to speak literally. So let us see if we have: if we have learned to speak literally then there should be no air of paradox about the following assertions:

There is no such thing as weighty evidence, there are no forceful arguments, I never find evidence compelling, no-one ever has any impressions of anything, nor does evidence ever sway anyone or incline them to one side or the other, evidence cannot tip the balance ... and so on.

All of these are metaphors. I suggest that remarks like, "there is no such thing as weighty evidence" is slightly shocking, and we are "inclined" to regard them as ("literally") false, rather than as the rejection of a metaphor. Once we notice these metaphors it also becomes evident how difficult (if not impossible) it is to avoid them.

We can trace the metaphor of the scales down the years to that most literal-minded of philosophers, Carnap. We can actually watch the metaphor going under for perhaps the last time, leaving behind only a single bubble to remind us of its continuing activity. In the course of Carnap's seemingly very non-figurative discussion of probability, we find this aside:

> In this point I am in agreement with Reichenbach, whose concept of weight corresponds to our concept of probability1. (Carnap [1962] p237 - 8)

and Carnap quotes Reichenbach:

> The man of practical life knows more about weights than many philosophers will admit (Carnap [1962], p238)

I take it that Carnap can only mean: "my literal-seeming discourse does the same work as Reichenbach's metaphor". In that case, the figurative words have been excised from the text, but not their figurative operations. If I do not like the metaphor of assaying evidence, then I may speak of assessing evidence instead: but I still mean "assaying", for I have not altered the way the discourse operates. I have only concealed the figurative operations in literal guise, or, to put it another way, I have altered the appearance of my figurative discourse to simulate "literalness". We might

equally well say that the "literal" discourse dissimulates its figurative operations. I have a hunch that Carnap's discussion of probability could be mapped onto a mathematical description of weighing something upon the weighing-scales.

It is also worth tracing the ongoing history of the scales of judgement. The weighing scales found their way into the chemical laboratory and (it was bound to happen sooner or later) chemical reactions came to be interpreted in terms of their equilibrium (an idea drawn from the operations of balancing, e.g. balancing weights upon the scales). Chemical equilibria, as I have argued previously, provide a metaphor for human judgement which is, in many ways, better than the "scales of judgement" metaphor. They have to be quite complicated chemical reactions to provide an apt metaphor for judgement. Some of the most complicated reactions take place in the human body and the human brain. However, if we push the metaphor very far in this direction, we are in danger of making the metaphor so realistic that it ceases to be a metaphor. Perhaps it is no coincidence that physical / electrochemical events within human beings should be complicated enough to match the complexities of the processes by which judgements occur. Perhaps the physical events within human bodies are the only processes complicated enough to embody processes of judgements. Then the development of the "chemical equilibrium" metaphor would be not so much a metaphor as a re-description (within a

different discipline and for different purposes) of the processes of judgement.

Templates ("innate ideas") are constructed. By invoking the metaphor of the template, we also make this constructive activity central to the processes by which we acquire beliefs.

Descartes aims to make the activity of construction even more salient. In the mnemonic systems developed since Simonides' time there was an active technology of memory: techniques were developed to connect ideas by "chaining" them together or hanging them upon "pegs". The place system of memory which Simonides used gave a way of ordering ideas, so that one could "find them back". The active part was making use of the techniques. Making use of this training was, in some ways, cumbersome (though not so cumbersome as being unable to remember at all). Lully's art of combinations was perhaps the leading edge of mnemonics by the time of Descartes: Descartes dismisses it in a single disparaging sentence:

> the art of Lully, for talking without judgement about matters one is ignorant of. (Descartes [1968], p20)

Descartes has much bigger ideas. As he writes, in a fragment which we have from Leibniz' copy of it (and the fact that Leibniz recorded it is not insignificant either):

> The sciences now have masks on them; if the masks were taken off they would appear supremely beautiful. On surveying the chain of the sciences one will regard them as not being more difficult to retain in one's mind than the number-series is.
>
> In the year 1620 I began to understand the foundations of a wonderful discovery. (Descartes [1970], p3)

This seems puzzling. For if we take just the natural numbers, for example, the series is infinite. How, then, could one retain them in one's mind? I take it that the point Descartes is making is that we do not need to retain the numbers, for we have a mathematical method for generating the numbers.

Descartes' wonderful discovery is a methodical technique for generating ideas - templates to fit the impressions upon the wax. For during the Renaissance, and leading up to Descartes' crucial role in developing the thought of the modern world, a central strand of discovery showed that we do not need to retain memories in the "wax" of the mind. For, building upon the Medieval rise of science, enquirers were learning to "read" what Galileo calls "the book of nature", and claiming that the book is written in the language of mathematics.

Descartes' guiding question (e.g. in the Meditations) is: what is the correct way to acquire true beliefs (and avoid falsehoods)? Translated into the metaphors I have mentioned, the question is answered as follows: we have to use a mathematical method to manufacture templates (ideas) which fit the impressions.

Descartes outlines the steps in the process - analysis into simple components, enumeration, ensuring completeness, and "orderly arrangement". How we do this orderly arrangement is a matter of free choice according to Descartes, but evidently he thinks that whichever way we do it, the result will be the same, for the example of orderly arrangement which he gives is the various techniques for finding all the anagrams which can be generated from a (finite) set of letters. Descartes' remarks warn us that these are phases of a process, for he mentions that enumeration, ensuring completeness, and orderly arrangement are all parts of one activity - i.e. they are not successive steps in the process - implying that other elements in his description are successive steps.

This is rather like making castings using the lost wax process: the simple elements are collected, enumerated, and put together in an ordered whole so that we can produce the template of the thing which caused those impressions.

So instead of having to retain all the impressions, we can regenerate them by using the templates (ideas).

The constructed idea has deep foundations (which fit the impressions), it is tightly constructed so that there are no gaps. But, we might object, there may be many ideas which "fit" the impressions. However, Descartes uses a mathematical model for the process of making templates, and the possibility of there being many ideas which "fit" would not show up in a model based upon the mathematics of his time. Mathematics has developed since: we now have the exemplars of (for example), alternative geometries (most famously used by Einstein) and, most recently, Chaos theory, describing processes which do not oscillate and do not tend towards stability (notice that the notion of "tending towards a limit" is crucial in the differential calculus developed by Leibniz and Newton, Descartes making a contribution towards that development). The usefulness of computability as a model for human understanding takes on a different meaning once we realise that, for some processes, a sufficiently powerful computer might be able to do the computation, but we cannot.

Gleich [1988] points out that the sort of mathematics available led practitioners of the physical sciences to concentrate on the solvable problems, leaving aside chaotic systems: the maths available led to force-fitting the data to the available

techniques, with the promise that other processes could be decribed by elaborating those techniques further - which was false. So that the physical sciences selected the data which suited them - namely those which fitted the mathematics available. Gleich adds that as a result we fooled ourselves into overlooking the existence of chaotic systems.

Descartes' method tells us that to find the theory that fits, we should start with the simplest one. If we adopt the (brave) assumption that we have a method for knowing which is the simplest theory, then the method can be performed mechanically, by feeding asssumptions into a theorem-generating machine and fitting the resulting predictions to the "impressions".

Descartes proposes that instead of recording impressions upon the wax, we need only to understand the construction techniques for beliefs: we can then construct, rather than retrieve from memory, any belief we require, as easily as mathematical procedures enable us to construct any member of a number series.

This at least was Descartes' dream. The metaphors explain the structure and sequence of Descartes' Meditations (Descartes [1968]). The sequence of the arguments is as follows.

1. The wax must exist in order for there to be any impressions (the mind is like wax) - this is the metaphorical structure of the cogito ("I think therefore I am").

2. The authenticity of the impressions is guaranteed by the impression of a template so perfect that it cannot be forged (the idea of God is so perfect that I cannot invent it, and the perfection needed to produce it guarantees that God, being perfect, is not a deceiver).

3. The authenticity of impressions means that they deceive me only if I confuse them - i.e. construct templates which order impressions in the wrong way, or if they are confused (one impression stamped over another one, perhaps) - or if the impressions are too faint for me to be able to ensure my template fits them (I cannot be deceived by ideas which are clear and distinct).

4. We can avoid error by not constructing templates which are not clear and distinct. If the templates do not exist then they cannot incline the scales of judgement towards falsehood. By confining ourselves to what is clear and distinct we ensure that the scales are only ever weighed down on one side, the side of truth.

5. There is a mathematical procedure for constructing templates: we reduce the materials to their simplest components, make sure that none are missing, and try to put them together in the simplest way. If they do not fit together this way, we try the next simplest way - until they fit together. The simplest one that fits (clearly and distinctly) is true.

Descartes' own text, however, is constructed not mathematically but metaphorically. If we consider the operations he must perform in order to construct the cogito, we can also explain why the connecting metaphor of the "template" disappears from Locke's model of cognition. Remember that the force driving both Descartes and Locke is the search for certainty: the old certainties have crumbled. They need a justification for a new order (and ordering) of things. The individual "subject" of experience has risen to prominence, and with it the problem of subjectivity. Descartes actually takes steps to construct this individual subjectivity: for he isolates himself in the famous "stove-heated room" of the Meditations, ensuring that the individual is thereby crystallised out of the flux of society. The isolated individual, rather than the social individual, becomes the unit of thinking. This isolating activity is needed to create the problems of subjectivity, for if thinking were to go on between individuals - e.g. in a conversation - rather than within one - as in soliloquoy - then a doubt as to

the existence of other individuals cannot arise: for they become the means by which I think and the subject of thought would not be Descartes' 'I' but 'we': instead of "I think, therefore I am" there would be, "we think, therefore we are". The problems would then be problems of intersubjectivity ("how do we know?" rather than subjectivity ("how do I know?"). Descartes constructs the conditions which make such doubt possible; but he could just as easily go back into society again, and thereby destroy his doubts.

Locke's way of overcoming the problem of subjectivity is to show that it allows objectivity: our ideas are aligned to external objects, provided we do not tamper with the "scales of judgement". Therefore a hint that our choices affect the processes of cognition is a threat to the project. The thought that we might construct the templates which link wax and scales is definitely a threat of that sort. It gives a major role to our constructive activity. Therefore the activity of constructing templates must be excised from the explanation.

Over the years, the models of rationality have developed. The Pre-Socratic notions of rationality were modelled on harmony and balance, the tension of the bowstring and the bow, the tuning of the lyre. Pythagoras, using mathematical techniques, found a measure of musical harmony. Next the model of rationality comes to encompass the precision of weighing and

measuring, one of its strands developing into the assessment (assaying) of probabilities - as found in Carnap's work.

I suggested that the aim of Simonides' "wax" metaphor was to enable human memory to be trained on the model of the best recording techniques then available - writing. With the development of photographic techniques it becomes possible to expand the optical metaphor - used by Descartes and taken over by Locke - and use it to model memory, the "retentive" quality of the mind, as well as its "receptive" qualities of perception. Photographs then offer a model for understanding, for they are so "natural" and "lifelike".

In the next chapter I shall discuss how Marx puts the optical metaphor to use in a way which subverts the aims which originally sustained it (namely the aims of establishing the veracity and certainty of one way of understanding, and thereby to justify it). After discussing Marx, I will be able to draw together the claims made in the last few chapters.

In this chapter I have tried to show how Nietzsche's "working hypothesis" may be applied to the history of one web of metaphors. I suggest that using the hypothesis enables us to understand much that would otherwise be puzzling. For example, Descartes' "proof" for the existence of God has seemed to many people an artificial intrusion into the main thrust of his

argument, but understood via the "wax" metaphor it becomes part of the pattern of the metaphor's operations. So that it is understandable even if we do not agree with it.

Nietzsche's working hypothesis allows us to propose ways in which understanding and misunderstanding may be constructed rather than passively received - through the "impulse to metaphor" and its radical interpretations which, far from being coerced or constrained by evidence, are the preconditions for our being able to construct evidence from data. Reading the texts in this way enables us to open them with a key.

It is remarkable that the same metaphors should keep surfacing over and over again, in the course of thousands of years. It seems rather too much of a coincidence. One explanation is that the metaphors are apt: they happen to fit (like a template) the non-figurative ways in which we discuss cognition, for example. Another option is to ask why they are so apt: might it not be the case that the metaphors fit because they guide cognitive discourse, giving rise to it in the way Nietzsche suggested? My suggestion is that the metaphors of wax, scales and template are only a small selection from a mutually-supporting system of metaphors which has grown, developed, and always guided our thinking. But there are other metaphors, and other ways of using the same metaphors.

The texts of Plato, Aristotle, Descartes and Locke are aimed to make something obvious. Suppose that we accept their proclaimed aim, to make the truth obvious. The same kind of metaphorical operations could go on in other texts, in order to make something obvious which is not true. For if we can construct evidence, (and we do, through the impulse to metaphor), we can choose what to make evident, i.e. what to make obvious.

We might then want to ask how we distinguish truth-seeking from self-deception. I suggest that we cannot distinguish self-deception by its characteristic operations: for they are precisely the same kinds of operations which are performed when seeking truth: the figurative construction of "obviousness", storytelling, "leaping" to conclusions", selective data-gathering, focussing attention, etc. Self-deception cannot be distinguished by the objects of attention: self-deceivers describe the same world as truth-seekers. If we are to distinguish the two, then we must pay attention to their aims and motivation. But someone who is self-deceived may believe his motive to be the disinterested search for truth - for he is self-deceived. How could a self-deceiver come to recognise that he was self-deceived? And if this is a predicament for the self-deceiver, it is just as much a predicament for "us truth-seekers". For our beliefs about our motives (disinterested search for truth, etc) will be exactly like a

self-deceiver's beliefs about his motives. So one wonders how the distinction is to be made. If it is not made, then the truth-seeker may find himself reflected in the activities of the self-deceiver. The result, for someone who wants to be sharply distinguished from self-deceivers (and don't we all) is paradox.

What was the guiding thought of Simonides, Plato, Aristotle, Descartes and Locke, if it was not to show that our way of understanding is justified, true, and certain because it is modelled upon the best technology we have available? And that if it does not live up to that technology, then it ought to do. There could be no better prelude to Marx's theory of ideology, the claim that we are deceived by participating in a mode of production which (a) requires us to adopt its instrumental theories in order to participate and (b) subjugates us to the products of our own labour and to the mode of production itself. Marx argues that ideology is deception, that we participate under duress, and that we can escape from that particular deception only by altering the mode of production - or at least altering the way we participate by struggling against it. To this I would add that, if Marx is correct, to participate willingly in that mode of production is self-deception. For in that case, one is not only deceived, but also responsible and willing to be deceived.

Ideology

My theory about self-deception applies to individuals. Marx's theory of ideology applies to whole societies and social classes. Writing on this large scale, Marx describes many features which I described on a small scale. Self-deception and ideology are closely connected. It is worth making the connection because the self-deceptions of individuals are not immune to the social circumstances within which they pursue their lives. The approach used by Marx to describe ideology underlines my claim that self-deception is (i) constructed by a process of interpretation and (ii) fixed by a role. The roles described by Marx are social (class) roles, the interpretations are those instrumental beliefs used by individuals to participate in a mode of production (such as capitalism). My theory, in turn, aligns with Marx's approach by illustrating how it may be used to dissolve the paradoxes of self-deception. Marx's theory offers a viable alternative to the false model of belief-formation which leads to the paradoxes, and so does mine. The two theories give each other mutual support.

To explain what Marx's theory of ideology contributes to the description of self-deception, I shall begin with some background information, then develop the connection.

1. Marx

> Men can be distinguished from animals by consciousness, by religion, or anything else you like. They themselves begin to distinguish themselves from animals as soon as they begin to produce their means of subsistence - their food, shelter, and clothing. (Marx [1977], p160)

Marx asserts that production is the basis of all societies. We need to produce and reproduce. We develop and alter the ways we do so, and these ways of organising production determine the way in which society is organised.

The form of a particular society is guided by the development of the forces of production (e.g. technological advances which increase our power to produce) and the relations of production between classes of people (e.g., in a capitalist society, the distribution of power between capitalists and workers).

Marx describes consciousness as a social product: it too is guided by the mode of production. An ideology is a way of understanding (or misunderstanding) which is guided by the needs of the mode of production within a specific society.

So, for example, in a capitalist society the need to accumulate capital governs the ideology. Human knowledge and

understanding are factors of production, and the factors of production are shaped and developed to serve the mode of production. So it makes sense to regard enquiry as a production process too, namely the production of human knowledge and understanding.

Capitalist production divides society into two classes, capitalists and workers; the relations between them are governed by the need to accumulate capital. Capitalists must attack the wages and conditions of workers, or be driven out of business and cease to be capitalists. Any capitalist who does not strive to maximise the exploitation of workers falls victim to others who do. For the capitalists who do so successfully will be able to control production at a lower cost. Capitalists are in competition with each other and with the workers who create capital.

To participate in the mode of production we have to adapt to the roles available within it. The roles adopted require us to act in ways determined by the needs of the mode of production. This also requires us to think in ways determined by the mode of production, for in order to perform those functions we need an instrumental understanding of the role. But the mode of production benefits capitalists at the expense of workers. Therefore, to participate, workers are obliged to adopt an instrumental understanding which damages their own interests.

This understanding need not be true in order to be effective. Under capitalism, Marx argues, it is false ("false consciousness"). Workers encounter a conflict between the ideology and their daily experience. They participate in order to gain benefits, yet by participation they enhance the power of capitalists who attack those benefits.

> The worker becomes poorer the richer is his production, the more it increases in power and scope ... the depreciation of the human world progresses in direct proportion to the increase in value of the world of things ... the object which labour produces ... confronts it as an alien being, as a power independent of the producer (Marx [1977], p78)

The situation can change. Competition between capitalists forces them to try to take more from workers - cutting pay and increasing working hours. As the rewards of participation diminish and the penalties increase, workers may cease to participate. By strikes, civil unrest, and ultimately by revolution they may struggle to overthrow capitalism. As a result their roles alter: they start to lose illusions which were fixed by the roles they formerly played as participants.

I suggested that self-deception is "ideology on a small scale". We can map Marx's remarks about ideology onto my remarks about

self-deception. An ideology breaks down when the rewards of participation in the mode of production diminish. Self-deception breaks down when the rewards of the role which fixes it diminish.

Lets take a more detailed look at what Marx says about ideology.

2. Marx's Theory Of Ideology

> If in all ideology men and their relations appear upside-down as in a camera obscura, this phenomenon arises as much from their historical life-process as the inversion of objects on the retina does from their physical life-process. (Marx [1977], p164)

> that is just the paradox of ideology: it is not just nonsense or error but "false understanding", a coherent, logical, rule-governed series of errors. This is the point Marx captures in his stress on ideology as a kind of optical inversion. In one sense, the inversion makes no difference at all; the illusion is perfect. Everything is in the proper relation to everything else. But from a contrary point of view the world is upside down, in chaos, revolution, mad with self-destructive contradictions. (Mitchell [1987], p172)

Ideology and self-deception are connected. One may participate to a greater or lesser extent in an ideology, and with a greater or lesser degree of willingness. An ideology has its agents - the professional ideologists - and its victims, the people who are deceived by it, willingly or not. The willing victims of ideology exemplify self-deception: it suits them to be deceived. The unwilling victims of ideology are, of course, deceived but not self-deceived.

Self-deception is like ideology, but practiced on a small scale: as well as the large-scale factors which may influence the thinking of whole classes of people, there are the small-scale factors which arise through the idiosyncratic circumstances of individuals.

Marx uses the camera obscura as a metaphor; but it does not originate with him. It had earlier been used for quite other purposes. Locke used the metaphor as a model of human understanding, to oppose the rationalist claim that ideas are innate or self-generated by the mind, by suggesting that ideas originate in the objective, material world "outside" the mind.

Marx, however, claims that,

> consciousness is ... from the beginning a social product [my emphasis]. (Marx [1977], p167)

Our understanding of the material world is not given direct but is mediated by social circumstances which we do not choose: "Man makes history, but not in circumstances of his own choosing."

Marx criticises Feuerbach for supposing that,

> the sensuous world around him is ... a thing given direct. (Marx [1977], p174)

Marx uses the camera obscura metaphor to assert that there is no such "direct" access. Our understanding is generated by the social machinery which constructs it. This machinery is developed by a person's "historical life-process". It is not guaranteed to construct true beliefs: it may (and does) construct false beliefs.

> When we stress the ultimate analogy of the physical eye [with the camera obscura - BC], we naturalise this machine and treat it as a scientific invention that simply mirrors the timeless, natural facts about vision. But suppose we reversed the stress, and thought of the eye as modelled on the machine? Then vision itself would have to be understood not as a simple, natural function to be understood by neutral, empirical laws of optics but as a mechanism subject to historical change. Vision would

> comprise not just the physiology of lenses and retinas but a whole field of ideological attentiveness - a preselected, preprogrammed grid of features and structures of perception. (Mitchell [1987], p175)

Allegedly the camera cannot lie. But what we see through the camera is culturally and historically conditioned. The camera mimics a particular way of seeing so well that it can be used to deceive. Yet because the artifice is so effective, the products of the camera are extolled as natural and lifelike.

Because the camera obscura mimics the "natural" way of seeing so effectively, the metaphor seems to support the claim that we have direct access to an objective, material world.

Marx draws attention to the fact that the camera is a product of a particular historical development of technology. If the camera is really so akin to understanding, then we may also consider understanding to be, like the camera, a product of a particular historical development of technology and a particular mode of production. The camera obscura is an artifice. It is not the guaranteed vehicle of truth. Likewise, human understanding is an artifice, moulded by social circumstances. Its veracity is not guaranteed: consciousness may be false consciousness. So we need to pay attention to the mechanism by which the appearance of veracity is produced:

> we do not set out from what men say, imagine, conceive, nor from men as narrated, thought of, imagined, conceived, in order to arrive at men in the flesh. We set out from real, active men, and on the basis of their real life-process we demonstrate the development of the ideological reflexes and echoes of this life-process. The phantoms formed in the human brain are also, necessarily, sublimates of their material life-process, which is empirically verifiable and bound to material premises. (Marx [1977], p164).

The connection of ideas to material life-processes is guaranteed for Marx by the means he chooses in order to explain ideology: but the veracity of the ideas is not guaranteed. Like the images of the camera obscura, everything may be in its proper relation to everything else though all are upside-down.

The ideology of a ruling class makes claims to universality and objectivity. Like Locke's camera obscura, it gives the appearance of being the (sole) "natural" way of understanding the world (misunderstanding it, in Marx's view).

Capitalismcreates ideologies. It also creates two social classes, the ruling (capitalist) class and the working class. Members of the ruling class dominate (by definition). Therefore they influence and to some extent control the life-

processes of the working class. Members of the working class use the available means to form their outlook. The dominance of the ruling class ensures that the means most readily available is the ruling class ideology. This way of understanding is shaped to serve the interests of the ruling class, with which the workers are, wittingly or unwittingly, in conflict. So when they look through this "camera obscura", their understanding is formed in a way which serves the interests of their enemies. By trying to use the ideology to further their interests, they actually damage their own interests. This can only be described as a misunderstanding. At best they may fail to "understand" in the manner made available. At worst they may obediently misunderstand. If they fail to "understand" then they are invited to accept the opinions of the experts who do understand. And those who do understand (or rather, "understand") are those whom the instrument is designed to serve. They "understand" with ease, because it is designed to serve their interests. The "understanding" prospers because it benefits the ruling class.

Marx's theory maps onto a Role theory of self-deception. It offers examples of how roles "fix" beliefs. Commodity fetishism (discussed below) is an example of such a belief. Marx also highlights the extreme difficulty of stepping out of some roles - it may take a revolution to do so in any consistent and lasting way - and so it casts doubt upon the

possibility of radical interpretation occurring independently of radical action (such as revolution).

Marx describes consciousness as a social product: it is not consciousness that determines being, but social being that determines consciousness. Human knowledge and understanding are factors of production, and the factors of production are shaped to serve the mode of production. So it makes sense to regard enquiry as a production process: in a capitalist society this process is controlled to produce forms of understanding which serve the interests of capitalists. So the dominant ideology is an instrument which is not well adapted to the needs of workers.

Suppose a member of the working class attempts to use the ideology. Some possible outcomes are listed below.

a. He may be able to alter his situation so as to become a capitalist. The ideology will then work to his advantage: he can use it instrumentally in order to understand (or misunderstand) the situation in a way which he finds rewarding. But this can only happen to a few individuals since it takes many workers to support one capitalist.

b. He will succeed in using the ideology as an instrument for understanding, without ceasing to be a member of the working

class. However by using it he misunderstands, and thwarts his own interests.

c. He will use the ideology, but because it is not well adapted to his needs, will be unable to make sense of the results. In order to function within the mode of production, he will be forced to accede to the wisdom of "experts" who do understand. These "experts" will be the agents of capitalism, for whom the ideology makes perfect sense because it serves their interests.

d. He will reject the ideology and try to construct another form of understanding. Since, according to Marx, consciousness is a product of the individual's historical life-process, this will be the understanding of a member of a defeated class subject to capitalism. The only effective way to create an understanding by and for the working class is to alter the historical life-process, through the self-emancipation of the working class. Unlike the ideology of the ruling class, this understanding will serve the needs of workers struggling to overthrow capitalism.

Capitalist ideology may be false yet still be an effective instrument. False understanding can be found in all sorts of societies of the past; many of these societies were very stable and, in their time, very powerful.

Marx gives many detailed descriptions of the workings of ideologies, in order to fulfil his promise to explain the development of ideology from the life-processes of human beings. In this chapter I can only give the flavour of what he says. One such illustration is his description of "commodity fetishism" (Marx [1977], p435 - 441).

3. "Commodity Fetishism"

Marx describes commodity fetishism as an illusion created by the capitalist mode of production. Commodity fetishism "reifies" the social relations between classes so that they are perceived indirectly, through their effects upon the market value of commodities. Social relations between people are misperceived as relations between commodities. Workers who fall victim to this misunderstanding are made less able to protect themselves in the conflict with the ruling class. For they do not recognise the real factor governing commodity prices, namely the conflict between social classes. Without recognition of the conflict, they are poorly situated to emancipate themselves.

The ruling class, in turn, are unable to avoid the courses of action dictated to them by the needs of competition, which is forced upon them by the logic of capitalist production. Capitalists must maximise the exploitation of workers in order

to accumulate capital, or fall victim to others who do so and who thereby control production at a lower cost to themselves.

Commodity fetishism asserts falsely that the value of a commodity is determined by its relation to other commodities, this value being measured by the workings of a free market. Marx argues against this that the values of commodities are determined by relations between people. These are the relations between producers of commodities, and the relations between social classes which are in conflict with each other. He calls the relations "social relations of production" for the mode of production creates and sustains them and they also influence the way in which production takes place. So, for example, capitalism divides people into the (large but hitherto defeated) working class and the (small but dominant) ruling class, the "agents of capital" or capitalists.

Capitalists can drive down the cost of production by making workers work harder, for longer hours, for lower wages. Workers can act to improve their wages and conditions, e.g through strikes. The balance of this struggle decides the cost of production, the rate of profit to the capitalist, and the market value of commodities.

Marx opposes to commodity fetishism a labour theory of value: the value of a commodity is the amount of socially necessary

labour time needed to produce it. Two remarks are in order here:

1. Clearly the labour time is not socially necessary if nobody wants the commodity.
2. The amount of labour time spent when using an outdated technology is not all socially necessary, for it could have been reduced using the more advanced technology available.

By participating in capitalism we are obliged to use the false theory (commodity fetishism), if only instrumentally. For if we participate, we are obliged to buy and sell commodities at the market price. Capitalist ideologies use commodity fetishism to explain that price. It prevents the price being decided by the labour time needed to produce the commodity, for capitalism alienates commodities from their producers. Therefore the price must be determined independently of the labour time. The result is a theory which cannot take account of the factors dictating the values assigned to commodities, i.e. the social relations between people. Nonetheless, commodity fetishism is instrumentally effective for the purpose of ensuring that capitalist production functions smoothly, to the benefit of capitalists and the detriment of workers.

To participate in the mode of production, we have to use the theory which makes its workings intelligable. If we want to

predict the price of a commodity in a capitalist market, for example, we had better not use Marx's labour theory of value, for it says nothing directly about the market price; and the data needed to apply the theory is not readily available, since capitalist ideology has no use for it. On the other hand, if we want to understand the social relations of production, we had better not use commodity fetishism, for it has little to say directly about those relations. However, commodity fetishism only appears to work because of the social relations of production. If those relations break down, because people cease to participate and actively oppose the mode of production - by revolution, for example - commodity fetishism breaks down and the real factors determining market values become apparent.

While the mode of production is functioning, commodity fetishism appears to work successfully. To make sense of capitalism in a way which allows one to participate, one has to use the false theory. If one used a true theory, one would understand that the ideology of capitalism is unintelligable, irrationality upon a grand scale, but one would not be equipped to participate. The result is a compromise. Workers have two sets of ideas, one set derived from the dominant ideology and, contradicting it, another set drawn from their own experience of class conflict. The balance between the two depends upon the balance of the struggle between the classes.

A capitalist ideology (of which commodity fetishism is a part) is shaped to make sense of capitalism, enabling one to participate in it. Marx offers a rival interpretation, designed to help workers to overthrow capitalism. This conflict between commodity fetishism and Marx's labour theory of value illustrates my earlier claim that an interpretation constructs order and disorder at the expense of rival interpretations, and also how a role may fix an interpretation.

If we use commodity fetishism without any mental reservations, then we are deceived. We may be unwilling, involuntary dupes, or it may suit us to be willing dupes of the theory - deceiving ourselves for the sake of the rewards of participation. If we have a mental reservation, it may nonetheless suit us to allow or encourage others to be deceived. If we do so, we are liars, colluding in the deception of others.

Another option is to oppose commodity fetishism. Marx suggests that this can only be done effectively by changing our practice, joining the struggle of the working class to emancipate itself. The shifting balance of the struggle between the workers and the ruling class governs the market value of commodities, including that commodity called labour.

Marx concludes that since (i) the working class has an incentive to alter its understanding, while the ruling class

does not, and (ii) consciousness is formed by the historical life-processes of individuals in society, therefore (iii) the only effective way to destroy the dominant ideological illusions is to change the historical life-process of the working class, by class struggle. So Marx's position is that a working-class ideology can only be developed through the struggle of that class for self-emancipation. It cannot be delivered to them by anyone else, for no-one else has the motivation to do so or the material circumstances which force them to struggle against capitalism.

Also, the situation (including the understanding) which would be created via that emancipation cannot be known in advance. For the historical life-processes needed to generate it do not yet exist. Therefore the reflective construction of "Utopias" is idle, and Marx condemns it.

We might suppose that Marx's criticism of capitalist ideologies is that they are insufficiently disinterested, but that is not quite it. Part of the criticism is that they claim to be disinterested when they are not. Part of the claim is that no ideology can be disinterested: it cannot exist unless it serves, or at least is compatible with, a mode of production. The criticism is that the interests represented are too narrow: we need a mode of production which does not betray people's interests. This mode of production will create, for those who

participate in it, an instrumental understanding which enables the mode of production to function.

> All previous historical movements were movements of minorities in the interest of minorities. The proletarian movement is the self-conscious independent movement of the immense majority in the interests of the immense majority. (Marx [1977], p230)

A working class ideology would represent a wider interest than a ruling class ideology: the widening of the sphere of interest would tend towards a limit, the classless (or one-class) society. There would be no other viewpoint from which to disagree with the (only) ideology of the (only) class. This limit is intersubjective, not objective.

If we had such a mode of production its ideology would be less misleading. Notice some of its features:

- the ideology would be overtly ideological, not naturalised: it would not conceal its status as a social product which could be revised

- the ideology would be recognised as a socially constructed instrument, designed to serve the interests of a wide range of people, not disguised as a "discovery". The aim of the

instrument would be recognised, not concealed, and it would serve the interests of all the people in the society

- human choices with regard to how something is to be understood would be recognised to be (a) possible and (b) legitimate.

These features are reminiscent of the radical interpreter's suggestion that we should retain our awareness of our own activity as "artistic creators" of interpretations.

Marxists, then, would reject the view that there is or ever could be an objective point of view - an understanding which is constructed without any of the machinery of a historical life-process, and which is disinterested, and which is independent of anything we do or think. We as historical creatures cannot attain such an understanding, so it is pointless (and self-deceptive) to aspire to it.

Ideology provides us with a model for many aspects of self-deception. Lets notice, in particular, the claim that ideology is not just nonsense or error, but false understanding: "everything is in its proper relation to everything else, but it is upside-down". This coincides precisely with the claim that self-deception is not merely a mistake nor merely a pretence but false understanding.

Marx's theory allows us to think about understanding and misunderstanding in a way which does not make self-deception paradoxical. It does not oblige us to respond to the paradoxes of self-deception by arguing that there is something wrong with the description, 'self-deceived'. It enables us to say instead that there is something wrong with models of understanding which make self-deception appear to be paradoxical.

Marx emphasises something which is not usually salient in discussions of self-deception. The form that self-deception takes will be strongly influenced by the social circumstances. The self-deceiver's interpretation is not going to be unrestricted and free, as Nietzsche sometimes appears to suggest. It is fixed by a role, and the roles available within society are not controlled by one individual. So one society will make some forms of self-deception much easier than would other societies. Different modes of production will encourage different beliefs. Some situations will make dishonesty and self-deception much more inviting than others. e.g. Mr Negligent relies upon his social position and some prevailing traditions to sustain his beliefs, and he reaps rewards for doing so. Mr Radical must work much harder to deceive himself, but has less incentive to be "negligently" self-deceived.

Self-deception, like ideology, will preserve the "proper relations between things, but turn them upside down". The

beliefs which observers think inconsistent will seem consistent to the self-deceiver. For a very simple reason. If we have no independent means of checking the "camera obscura", then upside-down is the only way we can perceive things to be: no other way of understanding can get in to disturb us. The only way of changing that situation is to amend the machinery which produces the understanding. This means a change to the historical life-process, with results which cannot be predicted in advance of the change. If one could predict the form of that different understanding, one would already have it, so it would not be a prediction of something in the future but an experience of something already present which generated it.

Self-deception, like ideology, is instrumentally effective (for the self-deceiver's purposes) without being true. It relies upon the processes which produce understanding (and misunderstanding), and the roles which fix the products of these processes by using them as beliefs.

Marx goes very far in the direction of the "process theory" of self-deception which I develop in a subsequent chapter of this thesis. For he (a) recognises the role of interpretation in the construction of "consciousness" (b) recognises that roles fix interpretations (for "consciousness is a social product" and participation in a mode of production such as capitalism causes "false consciousness"), and (c) he pays attention to the

processes by which consciousness is formed - the "historical life-processes" of individuals. Marxist analyses of these life-processes could be regarded as detailing in their minute particulars instances of the processes which I identify in the Process theory. My process theory, in turn, can be regarded as corroborating Marx's approach by showing how it enables us to dissolve the alleged paradoxes of self-deception.

In the next chapter I offer a model for understanding and misunderstanding - the "manufacturing process" model - which does not generate paradoxes when applied to instances of self-deception. This model is intended to link Marx's theory of ideology with Role theories and Radical Interpretation theories of self-deception. Following that, I shall give an example of how the model may be used to describe an instance of self-deception.

The Production Process

To understand self-deception we need to take note of the processes by which beliefs are generated. I do so now. Firstly, let us name this process. I call it interpretation. Enquiry is one special variety of interpretation.

I also want to displace the models of understanding which have been extant historically - the optical metaphor, the "scales of judgement" model and its associated metaphors such as "weight" of evidence, "force" of arguments, and so on. To do so I take manufacturing processes as a model. These processes have been extensively studied. We can distinguish several elements in such processes:

- raw materials
- resources (time, energy, etc)
- tools
- finished (or partly-finished) products
- by-products.

A manufacturing process may have several phases. We can map the elements of enquiry onto the manufacturing process:

raw materials = data

resources = time, energy, etc

tools = theories / hypotheses

finished products = evidence, beliefs

by-products = moods, emotions (perhaps - lets see as we continue this discussion).

The raw materials may be the finished (or partly-finished) products of other processes. So, for example, when a court of law engages in interpretation, some of the raw materials it takes in may be the testimony of witnesses - the products of the witnesses' processes of interpretation.

The tools - theories, for example - may be very general or designed for a specific task. Scientific theories are intended to be universal, applying to every specific instance. Bohr's theory of the atom, for example, is intended to apply to the constituents of every single table, tree, human being, planet, and so on. A theory about the causes of monetary inflation in the Middle Ages is much more specific; and a theory about why King Canute ordered the waves to retreat is more specific still. As a rule, interpretations use both general and specific tools. A parallel case in manufacturing would be, perhaps, a bottling plant: this includes both very specific tools (for bottling) and very general tools (the levers, screwdrivers, hammers etc used to maintain the plant).

We have a choice as to which of the things produced we count as finished products, and which we count as by-products. In the "mapping" above I treated evidence and beliefs as finished products; this is because for my current purposes I am interested particularly in beliefs (with a view to broadening our ideas about self-deception). Suppose, though, that a self-deceiver were more interested in sustaining a particular mood or emotion: then he might count beliefs and evidence as by-products, the primary aim of the process being to manufacture (or continue manufacturing) a mood or emotion.

I have mapped enquiry onto the manufacturing process. Interpretation is a broader kind of process: the finished products may be actions rather than evidence or beliefs, for example. Also, a manufacturing process does not occur in isolation from many other manufaturing processes: they will all interact within an economy. The nature of the process and the techniques used will reflect the development of that economy. And the economy will interact with other economies. Similarly, beliefs are not generated in isolation from the production of other beliefs. The processes which generate beliefs are not independent of the culture in which the person lives.

There are other factors which contribute to a manufacturing process. The process will require a good deal of organisation

and design; there must be a market for its products; the motivation must exist to perform the process; and so on.

There are various ways in which a manufacturing process may go wrong. There may be a lack of raw materials. This is like not having enough data to form an opinion.

There may be inadequate resources. This is like being too rushed, or too tired, to think about the data.

There may be insufficient or inadequate tools. This is comparable to those situations in the history of science where mathematics was not sufficiently developed to enable a particular understanding of the physical universe to be developed, or to the situation of people in the Middle Ages who did not have a sufficiently well-developed economic theory to identify the causes of monetary inflation.

The manufacturing process may be halted by an inability to remove the by-products (e.g. the storage space becomes choked with rubbish so that there is nowhere to put the finished products). A parallel case in the process of interpretation might be some kinds of disaster planning: if the contemplation of some kinds of catastrophe fills us with overwhelming emotions then we may be unable to assess the data in a way which suggests those catastrophes may be the outcome of our

situation. Conversely, if we are able to contemplate some situations without emotion then perhaps we are not assessing them properly either.

There may be no market for some finished products: they may be "neither use nor ornament". They may not be capable of performing the function for which they were intended (false theories; meaningless theories; inconsistent theories - like unworkable or self-defeating instruments).

The process may go wrong through lack of motivation: the production line is sabotaged or goes slow, or the goods produced are faulty, because the motivation is lacking. This is like the "epistemic negligence" I mentioned in an earlier chapter.

The manufacturing process may also go wrong because it is out of kilter with the rest of the economy. This is like the situation where someone acquires sets of inconsistent beliefs, or has theories which are incommensurable: incommensurable theories do not contradict each other (not overtly, at least) but they cannot be made to work in conjunction with each other.

There may also be too much motivation, so that resources are used to overproduce one kind of item, starving the rest of the economy so that other items are not produced. This is like

"obsessive" concern with a problem - or some details of a problem - which overshadows consideration of a broader view.

The process may also go wrong because the wrong kinds of items are manufactured. Even if there is a market for them, they may be destructive.

The design of the process may also be inadequate; if no design takes place, the process is likely to be incoherent too.

Having sketched in how the manufacturing model is supposed to apply to interpretation, let us see how self-deception fits in with the model.

Firstly, what about the claim that "the self-deceiver must know what he is up to?"

There are two common reasons for making this claim:

- the self-deceiver must know about the strategy of self-deception in order to guide and control it

- the self-deceiver must know that he is deceived because of the amount of counter-evidence - evidence, that is, which goes against the deception.

With regard to the counter-evidence, this is not available to the self-deceiver unless he performs some process to create the evidence. If he does not perform such a process, "the" evidence is not available to persuade him.

What he does have is the data from which the evidence may be constructed. But data is not evidence e.g. I may have knowledge by acquaintance with a rock, but that does not imply that I have knowledge about the rock - e.g. that it is 20,000 years old: to acquire that sort of knowledge I would have to do some interpreting.

Knowledge about the strategy used to deceive oneself is not available unless the self-deceiver performs some process to acquire that knowledge. This process will be additional to the self-deceptive process it is to interpret. The self-deceiver may not perform this additional process. He may not reflect upon what he is up to. If he does so reflect, he may not construe it in the way that our interlocutor would like: he may not use the theory our interlocutor does, and so not create the conclusions which the interlocutor wishes to draw.

A process which fails through lack of raw materials would be akin to ignorance (no data available); but if the lack of raw materials is due to sabotage, lack of motivation, etc, then it is more like wilful ignorance - self-deception by avoidance or

epistemic negligence. Similar considerations apply to a lack of tools and resources: it may be due to "sabotage".

If the market requires truth, then that is what will sustain the manufacturing process. Other market desiderata may be a desire for some kinds of comfort, or success.

Someone who is deceived may find out that what he believes is not true, when the product fails, perhaps in a spectacular manner. but what if (for his purposes, which may not include truth as a desideratum) the product does not fail? There is no reason why the process should not continue, even if it is self-deceptive.

Circustances are not always inimical to self-deception. One stock example which regularly appears in discussions of self-deception is the cuckold who is deceived about his wife's adultery and who wants to be deceived about it (Karenin, the deceived and self-deceived husband of Tolstoy's Anna Karenina, is an example). Often circumstances will work in his favour: others will be more than willing to deceive him and he will go along with being deceived. The product of this manufacturing process will fail only if the social support given by others also fails, or if the motivation for it fails - in which case the "market" for the belief collapses.

The aim of the process is not always to produce a belief. Sometimes the beliefs are by-products rather than primary products. Here is an example.

Suppose that many years ago Smith did me an injustice. As a result, I still harbour feelings of hatred and resentment against him. I would like to hit him, but he has since died. I also do not want to feel resentment against someone who is now dead. However Jones, my neighbour, resembles Smith; I notice that he has many mannerisms like Smith's. I pick a quarrel with Jones and hit him. I tell myself, "he deserved it: he's just like Smith".

I am not claiming that this process is rational. In the example, I had my reasons, but they were not good reasons. The aim of the process, actually, was to hit Smith; this aim was not fulfilled: instead I hit Jones who, in my mind, was a representative of Smith. What happened? I formed an interpretation, using Smith as a model for understanding (or, probably, misunderstanding) Jones. Smith's characteristics were mapped onto Jones; firstly his mannerisms - which fit, and so "justify" the rest of the process - then the things which I resent - which may or may not fit. The belief that "he's just like Smith", was a by-product of the interpretation. This by-product also catalyses the remainder of the process: it explains why, "he deserves it". It provides the rationale

for hitting Jones. I succeed in venting my resentment, for the time being; but since the ultimate aim of the process is not satisfied, it is quite likely to happen again, in an equally unsatisfactory manner. The aim of the process was not to form the belief: the aim was to express my resentment by some means, justified or not. I could not express my resentment without the belief, or something like it - except perhaps by simply "going berserk", "lashing out" - in which case I am taking the universe and everyone in it as representatives of Smith.

Also, I need not be conscious of the belief, nor of the process of interpretation. Perhaps I find myself "falling into" an argument with Jones, and simply know that I am going to hit him. Or perhaps I curtail the process: I manage to resist hitting him; I wonder why I resent him so much, and so on. In this case the self-deception takes place but it is not carried through to achieve its goal.

When the process is not aimed at belief, we may acquire beliefs as by-products without ever considering whether they are true or not.

How would the model of understanding used by Locke (for instance) handle a case like this? The answer is, I think, that it cannot. Locke provides a model of understanding, not

of misunderstanding. He makes some moves in this direction: he suggests that one can weigh evidence against the prospect of gain, for instance; and there is the suggestion of "tampering with the evidence"; but that does not capture what is going on here. In the example, 'I' was not tampering with evidence but constructing it. The choice of interpretation is explained by emotions (frustrated rage), desires (for revenge) and motives (to vent my frustration and rage). These did not alter the evidence. They were motives for constructing the evidence.

It may be objected that the model of the manufacturing process is so general that it is meaningless: what does it exclude? My reply is: what ought it to exclude? Human understanding can be very flexible and adaptable. Actually the manufacturing model does exclude quite a lot: there will be constraints upon understanding just as there are constraints upon manufacturing; and the constraints are just where we would expect them to be, for example lack of raw materials (data); lack of tools (theories) or inadequate tools (inadequate theories); inadequate design (negligence), competition (other processes give results better suited to the market), lack of motivation (results in shoddy goods and shoddy beliefs), etc.

A process which is self-deceptive will generate products (beliefs) which do not match the criteria for truth (whatever they may be) or which match them only accidentally: the

products will however match other criteria, and the beliefs may serve the aim of the process precisely because they are false, in circumstances where truths would thwart the process.

The Process theory of self-deception may look like rather a "thin" account of self-deception. Having asked how self-deception occurs, I have answered, "by a process". This may seem to be just another "magic button" of the kind I criticised in earlier chapters.

There is more to it than that. Self-deception is not any old process. It is a process of interpretation, and this chapter has modelled that process upon manufacturing. Consider how "thin" the rival interpretation is. It offers, for example, the "scales of judgement" model, asserting that there is an input (evidence) and an output (conclusions) with a process of "weighing" (unexplained) in between, the conclusions being determined by the "weight" of evidence (also unexplained). In the manufacturing model evidence is a product, not an input. The form it takes is related to all the inputs (not just data) as well as the "economy", that inter-related set of interpretations in which the person is engaged. It allows us to explain the influence of "background assumptions" without supposing that they are all dumped onto the weighing-scales of judgement.

The manufacturing model fits self-deception into a model of enquiry which matches the descriptions given by people engaged in enquiry (such as Szent-Gyorgyi's assertion that "discovery consists of looking at the same thing as everyone else, and thinking something different"). The alternative is an unworkable model of discovery in which evidence "compels" us to draw conclusions, yet some people are "compelled" to discover new conclusions while others are "compelled" to draw old conclusions.

Data (not evidence) is a crucial input to enquiry but, as Marx's labour theory of value makes abundantly clear, raw materials (data) are worthless until someone expends labour upon them. This is what the Process theory captures. Data do not constrain interpretation any more than apple trees constrain us to make cider. Data leaves us free to do what we like with them. But we are constrained in other ways.

Radical interpretation is a struggle against circumstances: against lack of tools (theories), lack of time, tiredness, lack of imagination, perhaps also mental anguish and the "economy" within which manufacturing occurs. This economy can assert itself when we try to create a new interpretation: a hostile psychic environment becomes apparent in one's own fears and prejudices, one's "sleepwalking" (as Koestler calls it), and trained incapacities to alter one's ways of thinking. The

hostility of other people may stifle discussion and the free flow of thought.

The misadventures which may befall a manufacturing process are not all occult psychic events. They occur in public, out in the open. The manufacturing model may be false, but it certainly is not empty. It makes a wealth of knowledge about processes available to be applied to the process of interpretation and to the particular case of self-deception. This knowledge is transferred from manufacturing. The metaphor is overt and unfamiliar. In my view this makes it better than the covert and familiar metaphors already "built into" our present model of enquiry.

A manufacturing process may be unsatisfactory in various ways. It may build ramshackle houses (compare the "foundational" theories of justification) or leaky rafts (compare the "coherence" theories of justification). The manufacturing model encompasses both of these ways of talking and makes them special instances of something more general. It acknowledges that jerrybuilders exist. They do not need any special strategies to pursue their odious activities. Sometimes they inhabit the ramshackle structures they have fabricated.

In earlier chapters I suggested that the "strategies" which allegedly characterise self-deception are not different to

those used in "normal" truth-seeking; rather, in self-deception the strategies which are used in truth-seeking are used with different motives. The manufacturing model captures this suggestion. Instead of regarding truth-seeking as a primary mode of understanding, and self-deception as a perversion of truth-seeking, we can now regard truth-seeking as one mode among others. This is surely correct: we do not gain the bulk of our beliefs by explicitly seeking out truths: we acquire them as a part (a crucial part) of all the other things we do, as the old saying suggests: "experience is what you get when you are looking for something else". We learn by trying to participate in the culture we are born into. We learn that to participate fully, we need to find out about things: so we ask, and, if we are in luck, gain an understanding (if we are not in luck then we gain a misunderstanding). Finding things out for ourselves comes later, when we have acquired some techniques for finding things out.

The model of the manufacturing process ties together the themes of interpretation and role-playing. The process of interpretation is guided by the role it performs: it links the raw materials (data), the market (the need for an interpretation which enables some specific goals to be achieved, such as guiding action to achieve a preferred outcome), the constraints of other processes (other interpretations) and the ways they use data and fit in with the

market). The products and by-products of the process are also inextricably linked together. So, for example, some interpretations "go with" some emotions and conflict with others. Using other models of understanding, there is little scope for overcoming the cognitive / affective divide, and the explanation of the link between beliefs and emotions is glaringly absent. The manufacturing model leaves room for this explanation.

What happens to the paradoxes, in this model? The self-deceiver is not inexplicably resistant to "the" evidence; rather, the self-deceiver generates evidence (and beliefs too) which is compatible with his other aims. The self-deceiver may or may not "know what he is up to". His understanding of the process may be prevented or modified by the existence of the process. It may be, like ideology, false understanding. Though false, it can still be operationally effective. Indeed, it probably is effective for his purposes precisely because it is false. The self-deceiver can be in control of the process, and responsible for it, without having true beliefs about it.

Let us use an example to see how this works. The example is taken from Cook [1987]. Cook, incidentally, offers it as an example of "deciding to believe without self-deception".

Nick is a creationist: he believes that the world was created by God in 4000BC. He also wishes to study biology at university, and he realises that, given the ascendency of the theory of natural selection, his creationist beliefs are going to cause him problems in advancing his chosen career in biology. He therefore decides to deceive himself into believing that the theory of natural selection is true. Six years later, he emerges from university triumphantly, laughing at the resolve he made to deceive himself: he had no need to deceive himself, for the theory he wanted to believe is so obviously true.

I want to extend this example a little, as follows. Whatever Nick may believe, his folks back home, who are also creationists, are horrified at what they see as Nick's shocking self-deception. Now take the story forward a few years. Nick retires, loaded with honours, at the end of a long and successful career as a biologist. He goes back to his roots, to his family; soon he realises that, after all, creationism is true. To the horror of his former colleagues, he repudiates the theory of natural selection. His former colleagues regret this development which they regard as a very salient example of self-deception. His family, however, regard him as having escaped from self-deception.

Notice how convenient it is for Nick, at each stage of his career, to have beliefs which enable him to participate and achieve what, in that environment, is regarded as success. It is tempting to suggest that, whichever theory is true (and they might both be false), Nick is a self-deceiver throughout the whole story.

So it is not obvious to me that Nick achieves his desired beliefs without self-deception. I accept that he has achieved beliefs at will. The fact that he regards them as obviously true does not alter my opinion, in fact it reinforces it: that is just what one would expect a self-deceiver to think. For how does Nick achieve his beliefs? I think the story might go as follows: suppose that Nick imagines (pretends) at the outset that the theory of natural selection is true, and then asks what else has to be the case. Starting from that point of view, he perceives that Creationism must be false. Why then should anyone believe it? There are motives for believing it: one may want the comforts of religion, one might want the freedom from doubt and the purity of purpose which can be available when one sticks rigidly to a spelled-out and extensive code of conduct, and so on. When one starts from the theory of natural selection, and then interprets everything else from that perspective, all "the evidence" must be consistent with it, because the evidence is generated by using the theory to interpret: the tools used shape the products of

interpretation. The forms of life which exist, and which the fossil record shows to have existed in the past, all fit in with the theory. This is hardly a surprise, for that is what the theory was designed to do.

Similarly if one starts from Creationism, and uses that to interpret, then the evidence constructed will be consistent with Creationism. Why would anyone believe in natural selection? It is a very powerful theory enabling one to make all sorts of correct predictions; it may free one from the restrictions and obligations which emerge from accepting a strict literal interpretation of the Bible. Some interpretations of data become very salient, for example the suggestion that random mutations of species could not have developed the forms of life now existing in the time alleged to have been available for the processes of natural selection to occur. This fits in very well with the thought that perhaps natural selection needs some assistance from a designer - a Creator. This interpretation may be a lot less salient in the mind of someone who accepts the mainstream view of natural selection. For he may regard it as a rather minor matter which will be resolved when our physical theories are better developed. I am offering no wagers about which one is right (and they might both be wrong).

One cannot put the evidence emerging from these two theories side-by-side in a single interpretation: one of them will always undermine the other. There is no neutral standpoint from which to interpret, since the evidence produced by one theory uses up the data which is needed by the other theory. One cannot gain access to the data in a theory-neutral way. Both interpretations use a theory (though not the same theory as each other) and so cannot be distanced from that theory. What one can do is to try out one way of interpreting and then the other. But to do either of them justice takes considerable time and effort, and one must approcah each of them with an "open mind" and some degree of patience, assuming that the things one does not yet understand will become clearer as one continues.

If one is not prepared to be "gullible", at least in the short term, then one will not have given the theory a fair chance: one will have simply judged it from the preconceptions and prejudices one already had. To give the theory a fair chance, rational criticism must be preceded by uncritical immersion.

There are dangers in uncritical immersion in a theory. Brainwashing (or "re-education") works by discouraging criticism and encouraging immersion.

Suppose, for example, that someone tells you that normal perception is inherently deceptive: to gain a true view of the world, one must take drugs which are "mind-expanding"; then one will be able to see the flying saucers which are invisible to normal perception and which really control everything. You take the drug, and you realise that the flying saucers really control everything. When the drug wears off, the realisation wears off as well. There is no neutral standpoint from which to choose between the perceptions one has as a result of ingesting the drug, and the perceptions one has after the drug has worn off. In your undrugged state, you can explain why such perceptions should emerge when one has ingested the drug; and soon after ingesting the drug, you can explain why the flying saucers, which are now so obvious, were invisible before. Without the drug, you can see that the drug inhibits your critical faculties and so prevents you making a proper judgement; with the drug, you can see that normal consciousness filters out lots of data which would otherwise undermine it.

The drug suspends disbelief in the flying saucer theory; critical consciousness suspends belief in it. The only reason scepticism has something to criticise is because we are already immersed in all sorts of theories which we learn when we learn our culture - through immersion.

Without immersion in a new theory, however, we "know" in advance that theories which rival our current theories must be false.

I think it rather unlikely that drugs alter specific beliefs in this way; but I think it very likely that drugs used within a cultural context can alter or enhance specific beliefs; and also that a cultural context can affect beliefs with or without drugs being used. The drug may make one more receptive to the cultural (or sub-cultural) messages. We participate in the culture by (partial or total) immersion. Thereby we gain an understanding (maybe a misunderstanding) of its practices (including the ways of talking).

The self-deceiver has an operationally effective understanding of what he is doing. This understanding is, however, when measured against the desiderata of truth (by "weighing the evidence"), a misunderstanding.

How do the other theories of self-deception fit in with the manufacturing model of understanding?

1. Schism theories.

There is no schism. If anything, the self-deceiver is all too well integrated. If there were a schism, then one part of the

self-deceiver would be free to criticise the self-deceiver's activity from a rival point of view.

2. Disconnection theories.

There is no disconnection. The self-deceiver's motives may well belie his beliefs: but the beliefs allow the self-deceiver to misinterpret the motives: the interpretation effaces its own motivation. Jean-Baptiste Clamence, the protagonist of Camus' novel The Fall, tells us that:

> I realized, as a result of delving into my memory, that modesty helped me to shine, humility to conquor, and virtue to oppress. I used to wage war through peaceful means and eventually used to achieve, through disinterested means, everything I desired. (p64)

With hindsight he recognises that his desire was to both dominate and to achieve popularity.

> The surface of all my virtues had a less imposing reverse side. (p64)

Yet at the time, he sincerely believes them to be virtues: he believes himself to be virtuous: the strategy effaces its own motivation. But is Clamence self-deceived, or merely mistaken

about himself? The answer depends upon the extent to which he is responsible for the strategy: if he is responsible, then he is responsible for being deceived about himself. If not, then he is simply mistaken. The strategy implies an understanding while its operation produces a different understanding - a misunderstanding. But this misunderstanding is not a mistake: it is a strategy. If it were a mistake, Clamence could not have pursued the strategy so unerringly to success. Clamence chose the strategy: he chose the mode of understanding (misunderstanding) in order to pursue his goals. Therefore he is responsible. Or, to put it the other way round: this misunderstanding arises from an underlying (instrumental, and not truth-regarding) understanding: the misunderstanding is the surface of the understanding. The misunderstanding is part of the strategy of someone who understands. Once again we have to note how well it suits the self-deceiver to misunderstand.

3. Role theories

The self-deceiver plays a role: he generates the role. The role is not merely a mask spread over the "real person within": the role is the real person, manufactured by the strategy of self-deception. The self-deceiver really misunderstands. Take the case of Clamence again:

> I have never been really sincere and enthusiastic except when I used to indulge in sports and, in the army, when I used to act in plays we put on for our own amusement. In both cases there was a rule of the game which was not serious but which we enjoyed taking as if it were. (p66)

By comparison:

> living among men without sharing their interests, I could not manage to believe in the commitments I made
>
> I lived my whole life under a double code, and my most serious acts were often the ones in which I was the least involved. Wasn't it this, after all, for which, on top of my blunders, I could not forgive myself, which made me react most violently against the judgement which I felt forming, in me and around me, and that forced me to seek an escape? (p66).

For Clamence, the role becomes unlivable:

> the engine began to have whims, inexplicable breakdowns. (p67)
>
> I pulled myself together, of course. What did one man's lie matter in the history of generations? (p67)

None the less the discomfort grew. (p68)

This is like the manufacturing process choked by its own by-products. For Clamence, the by-products include actions which do not fit in with his misunderstanding of himself as a man of virtue, and emotions (spite, rage, resentment) which are inappropriate to the role he has adopted. The misunderstanding ceases to be operationally effective: Clamence starts to experience things in a way which is not controlled by the process of self-deception: the interpretation sets off its own refutation:

> compliments became more and more unbearable to me. It seemed to me that the falsehood increased with them so inordinately that never again could I put myself right.
>
> A day came when I could bear it no longer (p69)
>
> I didn't want their esteem because it wasn't general, and how could it be general when I didn't share in it? Hence it was better to cover everything, judgement and esteem, with a cloak of ridicule. I had to liberate at all costs the feeling that was stifling me (p70)

Role theories relate to the design of a manufacturing process, and to the market requirements. The manufacturing process is

fixed by finding a market. Without a market it cannot survive. This is comparable to Marx's claim that ideology arises from the requirements of a mode of production.

4. Negligence Theories

Negligence theories relate to the market requirements of a manufacturing process: e.g. they criticise the self-deceiver for "falling away from the proper procedures for gaining truth": a sort of lack of epistemic quality control. They criticise the self-deceiver for being insufficiently oriented towards the truth.

5. No-such-thing theories.

These theories gain credence from the claim that self-deception is paradoxical. Since I have provided a theory of self-deception which is not (so far as I am aware) paradoxical, I give these theories no credence.

6. Radical Interpretation Theories.

These theories relate to the tools and techniques used in the manufacturing process, and to the prior manufaturing process which constructs them. 'Interpretation' names the process. Nietzsche tells us that the technique used is the artistic

creation of metaphors ('metaphor' being his label for figuration in general). Metaphors are tools which can be put to all sorts of uses. The design of the process, aimed at a market, fixes and spells out how the metaphors are to be used. In doing so it "literalises" the metaphors by drawing consequences - just as one can spell out the meaning of a sentence by listing the inferences which can be drawn from it.

The manufacturing process constructs order. By using raw materials and resources it also prevents them being available for use in other processes.

The manufacturing process model also gives us a way of explaining the differences between various kinds of enquiry. For example, a modern scientific theory is constructed and developed by an immense collaborative effort by many individuals, whereas philosophical enquiries tend to be pursued by individuals without much collaboration. Very few philosophical essays are produced by a team rather than by an individual, for example. Why is this? Questions cease to be philosophical when a methodology is established for answering them. The methodology forms the basis for collaborative work. Without it there is no agreed convention and so individuals, of necessity, "go their own way". Philosophical schools and traditions mark the early development of a methodology. Whereas a scientist who goes his or her own way will be unable

to participate in the scientific community, will be deprived of funding and collaboration. This bears out Nietzsche's remark that metaphors become "truths" through agreements which conventionalise them, and the conventions are needed as a basis for a social and gregarious way of life. When scientific theories break down, the basis for collaboration is lost, and scientists' activity tends towards metaphor-making and becomes more like philosophy ("natural philosophy"). Philosophers tend to shun metaphor when they are aiming to achieve agreement. Verificationism is perhaps the most readily identifiable "school" or "movement" in twentieth-century philosophy, and it is renowned for its dislike of figurative language. This dislike marks the aim to enable collaboration by sticking to the literal, spelled-out meanings, i.e. to not disturb the metaphors which are most well-established.

To conclude this chapter I summarise a defence of the Process theory.

1. The Process theory incorporates elements of Radical Interpretation theories Role theories and Negligence theories in a way which allows us to explain the attraction of the other (Avoidance) theories of self-deception - namely Dissociation theories and Schism theories.

2. The Process theory explains self-deception by putting it in the context of the "normal" processes of enquiry.

3. It lists the "factors of production" which are linked together in the process of enquiry - the inputs, the tools used, the "market" for its products, and so on. By connecting these factors it allows us to characterise the varieties of self-deception and to contrast and compare them with related activities such as wishful thinking, mistakes, lies, pretences and those mysterious failures of will which are labelled "akrasia".

4. The Process theory emphasises the role of "will" and responsibility in enquiry - aspects of the process which distinguish some deceptions as self-deceptions.

5. The Process theory disarms the "must know" problems which give rise to the paradoxes of self-deception. It leaves no glaring gaps about "what happens instead" of the "normal" ways of doing enquiry.

6. A remaining problem is the question of how are we able to assign responsibility to someone for being deceived - and therefore, how we are able to argue that he or she is self-deceived. The self-deceiver is "deemed" to be responsible; but we do not know if that is true or not. What we can do -

and the Process theory enables us to do it quite systematically - is to build up a case for someone being responsible, by reference to his use of the available "factors of production".

7. The Process theory provides a situation in which two philosophical traditions can meet and communicate. These are the "Anglo-Saxon" tradition (critical, analytical, literal-minded, rigorous, with an emphasis upon justification) and the "Continental" tradition (imaginative, figurative, gymnastic, with an emphasis on creativity). These are stereotypes but, I fear, easily recognisable ones. They also, interestingly enough, mirror the work done by researchers such as Sperry [1961] on the functions of the left brain and the right brain. It is worthwhile to persuade them to communicate with each other. For otherwise, both are impoverished and disabled.

8. The theory does not refer to "occult" psychologistic entities or processes, e.g. "unconscious intentions", which may be inferred but cannot be observed. All the processes to which the theory refers can occur in a publicly observable way. For example the creative construction of metaphors can be observed in brainstorming sessions. These are not commonplace in the contemporary practice of philosophy; perhaps they should be.

The thesis is not, therefore, an attempt to do "armchair psychology"; but it may be objected that the thesis is not

psychological enough. Research in other disciplines can solve puzzles in philosophy, so why insulate this thesis from the influence of psychology? I do not disagree: there is room for fruitful cross-fertilisation between the two disciplines. However, I do not claim to be qualified to carry out a research programme in the field of psychology, so it is pointless for me to attempt it. For my aims, the description of "unconscious" entities is superfluous in the description of self-deception, as Silver, Sabini and Macini [1989] makes clear:

> For the unconscious processing of information to be self-deception it must meet exactly the same criteria that the conscious processing of information must meet to constitute self-deception ... we suspect that the reason psychologists believe that unconscious processing is key was not that they discovered that unconscious processing played a prominent role, but because they believed that unconscious processing was conceptually required. Once we abandon this belief, we can probably dispense with unconscious processing as a constituent of self-deception altogether. (p223)

9. The Process theory places the issues of self-deception in a broader context which allows us to open other questions about the various forms of enquiry (such as: "why do scientists tend to work in teams while philosophers tend not to do so?").

10. The Process theory raises challenges to some long-standing epistemic theories. It disrupts the metaphors built into the traditions of epistemology ("weight" of evidence, etc) and puts in question old-established theories about justification (foundationalist theories, coherence theories). It favours the "pragmatic" claim that, far from commencing with a sound basis on which to build knowledge, we have to start from wherever we are, which forces us to make a "leap" (justified, if at all, by its results: "the proof of the pudding is in the eating"). Reasoning is a secondary, critical, evaluative process which cannot commence until an (unjustified) leap has been made.

This does not mean that reasoning has no role to play; but it does explain why reasoning can be blocked by a choice of interpretation, and why "weaving" between interpretations is beneficial. Reasoning at its best works in harness with imagination / figuration. The Process theory emphasises the importance of "artistic creativity" in philosophy and in enquiry generally, without demoting the "literal-mindedness" which views metaphors with suspicion, insists on spelling them out and analysing them, and so on. My regard for that activity is exemplified in my practice of trying to spell out the "magic buttons" used in explanations of self-deception. However, we need something to spell out, that is why we are forced to "leap".

Case Study: "Father And Son"

It is time to try out my account of self-deception on an example. I have chosen Gosse [1964] because:
(a) it is biographical and autobiographical (subtitled, "A Study Of Two Temperaments"), not fiction, (b) it describes an alleged case of self-deception in fairly extensive detail, (c) it describes (alleged) self-deception in a person whom the author knows intimately over a long period: his father.

I shall explain why these points are important.

(a) Fictional examples of self-deception may be "life-like", but that only means they appeal to our intuitions about what"life-like" fiction should do. Like fiction, biography is interpretation, and not guaranteed to be true. But, we hope, it is at least intended to be true, which we do not demand of fiction.

(b) It helps to have plenty of detail. A very sketchy description allows us to explain away the alleged self-deception ("perhaps he is just mistaken", "perhaps he is just

pretending", etc). The details explain why 'self-deceived' was the preferred description.

(c) Intimate knowledge of his father gives the son a wealth of background information about goals, motives and interests: we have as good an insight as possible into "what makes him tick". This is about the best we can obtain in the way of an example drawn from life.

To give the flavour of the book, here are several quotations:

> This was the great moment in the history of thought when the theory of the mutability of species was preparing to throw a flood of light upon all departments of human speculation and action. (p65)

> So, through my father's brain, in that year of crisis, 1857, there rushed two kinds of thought, each absorbing, each convincing, yet totally irreconcilable. There is a peculiar agony in the paradox that truth has two forms, each of them indisputable, yet each antagonistic to the other. It was this discovery, that there were two theories of physical life, each of which was true, but the truth of each incompatible with the other, which shook the spirit of my father with perturbation. It was not, really, a paradox, it was a fallacy, if he could only have

> known it, but he allowed the turbid volume of superstition to drown the delicate stream of reason. He took one step in the service of truth, and then he drew back in an agony, and accepted the servitude of error. (p65)
>
> The famous Vestiges of Creation had been supplying a sugar-and-water panacea for those who could not escape the trend of evidence, and who yet clung to revelation. (p65)
>
> Let it be admitted at once, mournful as the admission is, that every instinct in his intelligence went out at first to greet the new light. It had hardly done so, when a recollection of the opening chapter of Genesis checked it at the outset. He consulted with Carpenter, a great investigator, but one who was as fully incapable as himself of remodelling his ideas with regard to the old, accepted hypotheses. They both determined, on various grounds, to have nothing to do with the terrible theory, but to hold steadily to the law of the fixity of species.(p66)

To avoid confusion between the two Gosses, I shall call the biographer "Gosse Junior" and his father, "Gosse Senior".

Any example we may choose poses a problem, which anyone who shares the views of Gosse Senior can readily point out: any

such example suffers the disadvantages of perspective: we have to suppose that what the narrative describes is true: we must suppose that Gosse Senior really is deceiving himself ("accepting the servitude of error") and not, in however fallible a manner, groping his way towards a truth.

The problem is not that the doctrine of the fixity of species may turn out to be true. Even if it is true, Gosse Senior could be deceiving himself into believing it. The problem is that the description of his intellectual processes depends upon an interpretation of his motives and upon allocating responsibility for those processes: "he allowed the turbid volume of superstition to drown the delicate stream of reason" etc (my emphasis). The interpretation may be false. Perhaps he did not allow anything: perhaps he was overwhelmed by the "turbid volume" and so was not responsible.

What we can say, though, is that if the interpretation of motives and the allocation of responsibility is correct, then Gosse Senior was deceiving himself. And upon that assumption, we can go on to see how my process theory acounts for the self-deception. Gosse Junior describes its several phases:

1. "every instinct in his intelligence went out at first to greet the new light". So this is not a case of culpable ignorance or avoidance: if self-deception is achieved by

avoidance, then Gosse Senior lost the chance of deceiving himself at the outset.

2. "It had hardly done so, when a recollection of the opening chapter of Genesis checked it at the outset". But by then, for the Avoidance theorist, the opportunity had been missed.

3. "There is a peculiar agony in the paradox that truth has two forms, each of them indisputable, yet each antagonistic to the other." Gosse Senior could not have undergone this agony if he had deceived himself by a schism, or by partitioning belief in each theory into a seperate role which he alternated from time to time. This is not to say that schism was not an option for Gosse Senior, only that it was not the sole option and, as it turns out, it was not the option he adopted.

4. "He consulted with Carpenter ... They both determined, on various grounds, to have nothing to do with the terrible theory, but to hold steadily to the law of the fixity of species". This looks like avoidance, but, as we shall see, Gosse Senior does not stick to his decision. Instead:

5. "My father, after long reflection, prepared a theory of his own." In other words, the two antagonistic truths are to be reconciled by means of an interpretation. The data is to be

reinterpreted so that the unacceptable evidence is not generated.

> It was, very briefly, that there had been no gradual modification the surface of the earth, or slow development of organic forms, but that when the catastrophic act of creation took place, the world presented, instantly, the structural appearance of a planet on which life had long existed. (p67)

Gosse Junior goes on to say:

> In truth, it was the logical and inevitable conclusion of accepting, literally, the doctrine of a sudden act of creation; it emphasised the fact that any breach in the circular course of nature could be conceived only on the assumption that the object created bore false witness to past processes, which had never taken place. For instance, Adam would certainly possess hair and teeth and bones which it must have taken many years to accomplish, yet he was created full-grown yesterday. (p67)

Charles Kingsley reacted like Gosse Junior: he could not,

> give up the painful and slow conclusion of five and twenty years' study of geology, and believe that God has written on the rocks one enormous and superfluous lie. (p68)

This is clearly not the proposal which Gosse Senior was making: he was not suggesting that God is misleading us, but that we can mislead ourselves. The way we do so is by not taking account of the truth revealed by the inspired writing of scripture.

I think that the remarks made by Kingsley exemplify part of the the misapprehension about evidence which I have pointed out already. Rocks do not mislead us. We may mislead ourselves with regard to them by misinterpretation. Rocks are data, not evidence: we create evidence by interpreting data. So I feel obliged to defend Gosse Senior on this one point. Gosse Senior's theory certainly does invite us to say that e.g. the fossil record is misleading; but equally, there are scientific theories which imply that all sorts of things are misleading, including the illusions so often referred to in works of philosophy - the stick in water which looks bent although it is really straight, the optical illusions used by psychologists to study perception, and so on.

Gosse Senior would be entitled to argue that the only people who are misled are those who choose to be misled, namely those

who reject a theory which explains why we are misled by the appearance. If we use a theory about fossils in order to say that some rocks existed before the creation of the universe (as dated by the Bible, interpreted by Gosse), then we are misled by applying a theory where it does not fit: applying a theory about the development of geological and organic systems to the moment of creation, indeed to an (impossible) time before the creation of the world.

This is the only point on which I wish to defend Gosse Senior. The example matches up with my criteria defining self-deception.

Firstly, the self-deceiver must not know what he is up to. Gosse Senior does know that he is attempting to reconcile two antagonistic theories. He also knows, in detail, what the two theories say. "It was not, really, a paradox, it was a fallacy, if he could only have known it". Gosse Junior is offering a plea of mitigation here: his father did not know what he was doing; "but he allowed the turbid stream of superstition to drown the delicate stream of reason". He did not know what he was doing, but he was <u>capable</u> of knowing. "He took one step in the service of truth, and then he drew back in an agony, and accepted the servitude of error". So what prevented him going further was "agony". His motive was the agony of rejecting a literal interpretation of scripture.

Notice how rational Gosse Senior's approach was. He was faced with a new theory. Its predictions (or retrodictions) could be extrapolated back into the past. But there is always a danger with such extrapolations: the further we take them, the more we risk extending them beyond their proper scope, and the more other factors may interfere. For example, on more than one occasion economists have failed to predict economic booms and slumps because they extrapolated current trends.

Gosse Senior is prepared to extrapolate the predictions back to the beginning of the universe; he also believes that he knows from an independent source (interpretation of scripture) that the universe is not infinitely old, that it had beginning, recorded in the Bible as the moment of the Creation. He therefore claims that the theory cannot be extrapolated back further than that.

Kingsley argues that Gosse Senior is thereby committed to a claim that "the rocks tell a lie". Gosse Senior can reply that we are not obliged to adopt an interpretation which misleads us. Indeed, the revelation contained in scripture (which Gosse Senior interprets literally) should prevent us being misled. We are only misled if we wilfully reject scripture. It would be a strange sort of lie if its author also gave us the means to avoid being misled by it.

Consider how a Negligence theorist would interpret this example. In what way does Gosse Senior "fall away from the proper procedures for gaining truth" or "fail to play the proper game of knowledge and ignorance"? Gosse Senior approaches the dilemma in a very rational way. Admittedly he does not (for example) go out and seek for more evidence. But given the state of his knowledge, it is not clear how much more evidence he needs, nor how, supposing he gained extra evidence, it could alter his situation.

Arguably he is at fault in keeping his literal interpretation of scripture immune from criticism. He does not subject it to further evaluation. But he believes that he has independent grounds for the interpretation: Kingsley never puts those grounds in question; and Gosse Senior has found a way to reconcile that interpretation with the theory of natural selection which supposedly contradicts it. There is no fault of logic in his resolution of the dilemma. His critics find it implausible. But "it seems implausible" is not a very good argument without some reasons to back up the impression of implausibility. The (inadequate) only reason offered is that his resolution makes God (or "the rocks") a liar.

What is going on is a clash of interpretations. The disagreement is not about the data: every side agrees on the data. They disagree upon the import of the data.

Suppose that Gosse Senior had drawn his restriction upon the extrapolation of natural selection not from Biblical literalism but from, for example, a theory of physics. I do not believe that he would have been dubbed a self-deceiver in the same way; for what is at issue is not what is alleged to be at issue, but the relative authority of different grounds for belief; and physics would have been granted high authority as one of the "hard" sciences; unless there was something extremely dubious about the theory from physics, his "resolution" would have been granted more respect. However Biblical literalism was not accorded the same respect: perhaps for good reasons; but these good reasons are not made apparent in remarks like Kingsley's.

My point is that the procedures followed by the self-deceiver do not differ strikingly from those of someone who is not deceived at all, or someone who is just mistaken. Gosse is following the proper procedures for gaining truth, but from a different perspective than his critics. Gosse is playing the proper game of knowledge and ignorance. Undoubtedly he has his motives which lead him to try some interpretations rather than others; so, I suspect, have his critics. Perhaps they are just luckier than he is, in that their motives happen to square nicely with the situation - i.e. they pursue the truth because it suits them to do so, just as he moves away from the truth because it suits him to do so. I do not see how this makes him

any more reprehensible than anyone else. He is doing exactly the same sort of things as they are, but starts out from a less fortunate position.

Schism theories and Dissociation theories have little chance of explaining this example. Gosse Junior, who knows his father intimately, argues that his father does not know the truth about which he is deceived. His father "draws back in agony" from that knowledge, knows of the theory of Natural Selection but does not believe it, and instead develops his own theory. Could Gosse Junior have misdescribed the situation? A Schism theory forces us to postulate a seperate agency within Gosse Senior, unknown to the deceived part of him, and also undetected by his son. We could argue that the son was deceived by the very fact that he was too close to his father, and therefore unwilling to detect a schism. But to do so seems pointless. The proposed seperate agency explains nothing that cannot be explained in other ways. Our only reason for proposing it is attachment to a Schism theory because (we allege) it is the only way to avoid the paradoxes of self-deception. But the allegation is false. We have a non-paradoxical and effective explanation which does not force us to postulate a schism.

Dissociation theories fare no better. The suggestion that Gosse Senior might have disconnected his belief in natural

selection from his actions, emotions, and other items of mentation, does not square with his son's description of an agonised and earnest struggle to reconcile scripture with the theory. This is not dissociation but agonised association.

The same reasons militate against our describing this case as the adoption of a dissimulating role. Gosse Senior aims to reconcile two theories, not to defer or mask a belief.

The example looks much more like radical interpretation. Gosse Junior describes the process of constructing an interpretation, which Gosse senior publishes in a book. This interpretation is fixed by a role (a role simulation): Gosse Senior uses it to interpret data and to guide his emotions (he is pleased with his interpretation, and dismayed when it is soundly rejected by public opinion). He also uses it to guide action, continuing both his religious activities and his scientific researches.

One last suggestion remains: perhaps this is not a case of self-deception at all. Perhaps Gosse Senior is merely pretending to believe or is simply mistaken. Yet the agonised, elaborate construction of an interpretation is not a simple mistake, indeed it seems not a mistake at all. Some of Gosse Junior's remarks make it seem like a mistake ("could he but have known it", for example); other remarks make it look like a deliberate pretence ("he drew back", he "accepted the

servitude of error") and sometimes both aspects appear together ("He consulted with Carpenter, a great investigator, but one who was as fully incapable as himself of remodelling his ideas with regard to the old, accepted hypotheses. They both determined, on various grounds, to have nothing to do with the terrible theory, but to hold steadily to the law of the fixity of species"). "Incapable" suggests a mistake, "they both determined" suggests a deliberate act and so not a mistake. However, neither description fits the example well: Gosse is not simply mistaken; he is not merely pretending. He can be described, in a way which is neither paradoxical nor misleading, as self-deceived.

Radical interpretation, role simulation, and the process theory of self-deception suffice to achieve this description. Gosse constructs his theory by a process, radical interpretation, and fixes it by role simulation. I have given an adequate description of an instance of self-deception. In this case, at least, my theory of self-deception works.

The line of enquiry favoured by Gosse Junior and Kingsley was, for Gosse Senior, blocked. It was agonising for him to take that route, so he looked for another way - and found it. Notice that if the agony were very great then we might decide not to call Gosse Senior a self-deceiver. For while there are some things we expect people to face up to, there is a limit to

what we expect: if someone cracks under relentless torture, we are unlikely to blame them. Whereas if someone cracks under the hint of some possible slight miscomfort, we may well blame them. Gosse Senior, I take it, was somewhere between these two extremes, and so somewhat responsible. He "allowed" superstition to overcome reason.

Instead of using the new theory, facing up to the agony and pursuing the truth (as his son saw it), Gosse Senior constructed a different interpretation. His various roles fixed the interpretation. His religious practices fixed his way of interpreting scripture, his scientific investigations fixed his commitment to the methodology and powerful means of classification which the new theory made available - and, his son tells us, Gosse Senior was above all a collector and classifier. His interpretation was closely tied by his emotions:

> he found it unthinkable that he should modify his belief in the literal truth of the Bible, even when scientific evidence seemed directly to contradict it ... In the longer term, it no doubt helped to explain his increasingly unreasonable religious fervour, which ultimately drove his son to "rebel". [publisher's note to 1964 edition, p.vi].

This summarises a process which Gosse Junior illustrates with a bookful of circumstantial detail. Gosse Senior was the self-appointed minister of a small community of fundamentalist believers. If he had been tempted to waver in his religious beliefs, this role must surely have made it more difficult to do so. He was also a scientific researcher with connections at the Royal Society and the British Museum. If he had been tempted to abandon the commitment to scientific method, this role must surely have made it more difficult. The two major themes of his life were in conflict, and by constructing his theory he tried to prevent them dashing each other to pieces; but in consequence,

> by a strange act of wilfulness [namely, constructing and publishing his theory], he closed the doors upon himself forever. (p66)

His public role as scientific enquirer was destroyed, though he continued his scientific investigations.

> He had been the spoiled darling of the public, the constant favourite of the press, and now ... he could not recover from amazement at having offended everybody by an enterprise which had been undertaken in the cause of universal reconciliation. (p68)

Gosse Senior needed his theory in order to sustain his two roles, and (putting it the other way round) his two roles demanded that he develop such a theory. He created it by interpretation, but the two roles fixed it, and made other options "unthinkable". This was not because the roles were comfortable pretences, but because they were major parts of his personal identity. The roles were no facade. They did not mask something more real: they were the reality.

Gosse Senior was a radical interpreter, and he was punished for it. Alongside him, I suggest there were many negligent self-deceivers, who went along with an interpretation because it was easy, because it suited them, and because their roles (different to those of Gosse, and lived out in different circumstances) fixed their beliefs in different, easier, and more socially acceptable ways. I suggest there were also plenty of mobile self-deceivers who moved from a fundamentalist interpretation of the Bible to the new theory of natural selection as it became established, as they recognised the opportunities and advantages associated with the roles which fixed the new theory.

Hybrid Theories

I have discussed the "pure strain" theories of self-deception. Now I should mention some representative "hybrids". The classification is tentative because in many cases I am not sure to what extent a writer's allusion to a "pure strain" explanation is explicit and intentional rather than implicit and unintentional. This is not surprising since the classification is mine rather than theirs and was created after they were writing. Some influential works have not been cited in the body of this text. I have listed them seperately in the bibliography, to assist others who embark upon a survey of this literature.

1. Negligence and Schism Theories Combined

Fingarette [1969] suggests that the unity of an integrated self is achieved by using an acquired skill, "spelling out", in order to make oneself aware of one's "engagements" - the projects one is engaged in. Self-deception occurs when someone neglects to spell out an "engagement" and therefore does not integrate it into consciousness. So I regard Fingarette's theory as a combination of schism (lack of integration) and negligence.

2. Negligence and Role Simulation Theories Combined

Plato combines a Negligence theory and a Role Simulation theory: self-deception is mis-integration rather than dis-integration since the self-deceiver by negligent "injustice" makes reason subservient to other elements of the soul, using reason to pursue the (unjust and unreasonable) role or, for example, an oligarch.

Peterman, criticising Avoidance theories, argues that negligence alone is sufficient to explain self-deception.

3. Negligence and Radical Interpretation Theories Combined

Butler's sermon on "self-deceit" suggests that negligence is a necessary condition for self-deception but not that it is sufficient: negligence allows moral vice to extend to the intellect, for without a proper examination of conscience one is almost inevitably deceived by "self-love". I suggest that Butler combines a Negligence theory with a Radical Interpretation theory (self-love alters the way in which one interprets one's actions).

Mele [1983] offers a Radical Interpretation theory:

> the following is a characteristic and jointly sufficient conditions of a central case of S's entering self-deception in acquiring the belief that p.
>
> (i) The belief that p which S acquires is false.
> (ii) S's desiring that p leads S to manipulate (i.e., to treat inappropriately) a datum or data relevant, or at least seemingly relevant, to the truth value of p.
> (iii) This manipulation is a cause of S's acquiring the belief that p.
> (iv) If, in the causal chain between desire and manipulation or in that between manipulation and belief-acquisition, there are any accidental intermediaries (links), or intermediaries intentionally introduced by another agent, these intermediaries do not make S (significantly) less responsible for acquiring the belief that p than he would otherwise have been. (Mele [1983], p370)

Condition (i) is incorrect: the belief may be true; (ii) is also incorrect, for the negligent self-deceiver may adopt the belief that p, not because he desires that p, but because the belief is an effective instrument for achieving his desired goal. Nonetheless, the core of Mele's argument is correct. Mele lists "some common ways in which a person's wanting that p may contribute self-deceptively to his believing that p":

"negative misinterpretation", "positive misinterpretation", "failure to focus", and "one-sided evidence-gathering". The terminology itself indicates that Mele's is a Radical Interpretation theory. Mele also emphasises (p371) that "the chief virtue" of C is that it is (in my terminology) not an Avoidance theory but a Not-Know theory.

Mele's treatment of self-deception is made easier by (i), and also by his implicit appeal to epistemic norms ("to treat inappropriately", quoted above, implies that there is also an appropriate way to treat data - a way which conforms to the norms). I have denied myself this, but, I claim, the result is both more enlightening and more disturbing.

Fingarette writes that the cognitive aspects of self-deception have been over-emphasised at the expense of the volition-action aspects, and that he is restoring the balance. Non-cognitive accounts of self-deception follow Fingarette's lead (e.g., Hamlyn [1971], Whisner [1983], Solomon [1978], de Sousa [1978]). De Sousa, following the remarks of Broad [1971], argues that emotions can be construed as cognitions or judgements. This is perhaps ironic in view of Fingarette's argument. De Sousa argues that to accept an unexamined emotional "ideology" is self-deceptive, and inauthentic:

> in what I have been calling "self-deceived" emotions the self usually connives rather than originates. We are responsible only to the extent that we are generally motivated to conform to the social and gender role assigned to us and that we allow ourselves to be taken in by the feigning this necessarily requires. (De Sousa [1978], p693)

De Sousa combines a negligence element with a radical interpretation element (indicated by his reference to an "ideology" of the emotions). The references to "conniving", "originating" and "feigning" may also indicate a process element.

4. Radical Interpretation and Role Simulation Theories Combined

I have traced one line of theories, from schism to radical interpretation. In sections 5 onwards I consider a second trail, which leads from Dissociation theories to Role Simulation theories. When we put the ends of the two trails together we find the hybrid theories put forward by Nietzsche and Marx: radical interpretation explains how one can perform role simulation, and role simulation explains how radical interpretations can be "fixed" as beliefs.

5. Dissociation and Negligence Theories Combined

Bach [1981] exemplifies Dissociation theories, arguing that the self-deceiver dissociates what he believes from what he "thinks". Self-deception is a feat of directed consciousness, with the self-deceiver focussing on some items while ignoring others. I agree with Hellman [1983], who criticises Bach:

> it is puzzling that one could think that not-p on a sustained, recurrent basis, [while believing that p].
>
> [Bach] is content ... with letting a kind of psychological causation (motivation) account for this phenomenon ... psychological processes and concepts are supposed philosophically unproblematical. Lest we forget, "self-deception" itself may be taken as a "psychological process": this has certainly not stopped philosophers (including Bach) from raising questions about it.

Pugmire [1969] finds the dissociation in a different place. Pugmire distinguishes between "the truth" and "the glaring truth". A self-deceiver may know the truth without ceasing to be deceived, but "the glaring truth" can destroy the deception. Hellman's questions are just as pertinent when applied to Pugmire: how could one perform this feat of sustained

attention or inattention to something so important that it drives one into self-deception?

The relation between the dissociated items was always conceptually peculiar in Dissociation theories. It performed the dual (not inconsistent) roles of linking the items while keeping them apart. The initial tenuous bridge between a belief and consciousness becomes, in Pugmire's account, a broad highway. "Dissociation" is giving way to quite a different account of self-deception, in which what matters is not what one believes but how one believes it - teetering on the brink of becoming a Role Dissimulation theory. If a self-deceiver is to achieve the precarious feat of directed attention, as required by Bach and Pugmire, then adopting a role is an obvious way to make the task easier (indeed by adopting a role one makes it much less precarious). Hamlyn [1971] is even closer to this brink, arguing that one may be self-deceived by witholding one's genuine emotions, without masking the true belief which gives reason for having those emotions. This can be characterised as a Dissociation theory (the self-deceiver dissociates what he believes from his emotions); or it can be characterised as a Role Dissimulation theory (the self-deceiver plays a role to evoke some chosen emotions despite their conflict (or the conflict we think there ought to be) with the beliefs.

Demos offers a Role Dissimulation theory. He points out that someone who intentionally pretends to himself is only behaving as if he believed, "make-believing" rather than genuinely believing, and he knows that he is doing so. Demos suggests that self-deception occurs when the self-pretender begins to be unwittingly taken in by his own performance and begins to "believe his own lies": role dissimulation becomes Role Simulation. The dissimulation is the voluntary element of self-deception; it paves the way (negligently?) to an involuntary outcome. Demos implies, without spelling out, something which is crucial to explaining self-deception: a process of radical interpretation ("self-pretence").

6. Dissociation and Role Theories Combined

Martin [1979b], commenting on Factor [1979], suggests that four features are typical of self-deceiving forms of self-pretence:

> (1) the pretence is engaged in with the purpose of evading a confrontation of an unpleasant reality that one is responsible for facing
>
> (2) although the person knows he is engaged in this evasion, he refuses to admit as much to himself, and this involves witholding his genuine emotions towards the reality he is fleeing.

(3) while the person knows that he is pretending to himself, he does not sincerely confess this to himself.

(4) in both (1) and (2) the avoidance of the self-admission does not require coming to believe the opposite of what he knows.

This epitomises Dissociation theories. They soften the "Avoidance" requirement, that the self-deceiver must both believe that p and believe that not-p, suggesting instead that the belief is disconnected from other items. In Martin's theory, the disconnection is between:

- knowing that he is engaged in an evasion, and admitting it to himself
- knowing that he is pretending to himself, and sincerely confessing this to himself
- knowing about the unpleasant reality that one is responsible for facing, and confronting that reality (Martin in one of his examples writes of an instance where, "his refusal to confess his emotions and pretence to himself is not a matter of belief, but rather of impeding and stifling his emotions" (p442).

Self-pretence certainly would seem to militate against anything so solemn as "making sincere admissions to oneself"; so it

could at least delay awareness of what one knows about oneself. And "delaying awareness" could fall within the scope of self-deception.

Martin [1979a] adds that self-pretence,

> can also be engaged in with the intention of fleeing a painful situation or aspect of oneself that it is one's proper business to explicitly recognise and deal with. Here make-believe is an exercise in hiding from oneself, whether or not it eventually leads to a self-deceiving belief.

7. Missing Hybrids

The diagram on the following page summarises the connections between the hybrids I have mentioned. Notice that there are some missing links. Other theories make schism theories redundant, so they rarely combine with schism theories. Dissociation theories, in particular, postulate a connection between the dissociated elements in the self-deceiver, so they actually deny that there is a schism. Radical Interpretation theories make Dissociation theories redundant, and so do not combine with them. Role Dissimulation explains how Dissociation occurs - and so makes it redundant. Role Simulation theories make it unnecessary to postulate a "masked"

element in the self-deceiver - making Dissociation theories redundant. No-Such-Thing theories, naturally, are inconsistent with all the other theories.

8. Kipp [1980] gives a No Such Thing theory, distinguishing "ameliorists" who depict self-deception as "ignorant mistakenness", "euphemists" who depict self-deception as "mere dishonest pretence" and "literalists" who depict the self-deceiver as someone who both does and does not believe something and who "fools" or "persuades" himself. Kipp argues that "self-deceivers",

> are trying to fend off, through deceptive pretence, what they regard as defeat, or unacceptable loss of face, in a not entirely unreal, socially-staged power struggle, or status-seeking contest, whose goal is to appear, in the eyes of others, a maximally enviable existential success. (Kipp [1980], p315)

> taking an uncritically charitable view of certain motives and behaviour is what most decisively misleads people into concluding that literal self-deception must be possible. (p317)

Kipp suggests that"self-deception" needs an audience. For example, a widowed recluse receives news that her son has been

killed, but "deceives herself" into believing that he has not. In the absence of an audience to witness this,

> the mother would have to be seen as distractedly grief-blind rather than as literally self-deceived. (p316)

I think that a more important consideration would be that if one cannot help being bereaved, one can at least avoid feeling bereaved. For this purpose, self-pretence may be sufficient.

Some hybrids are absent from the diagram. Role Simulation and Radical Interpretation theories make other "pure strains" redundant and so do not need to be combined with them. Schism theories have few hybrids for the opposite reason: the other theories make Schism theories redundant. No-Such-Thing theories, naturally, do not combine with Such-Thing theories (i.e. all the other theories). Dissociation theories have lots of adherents and lots of hybrids. The reason is that (a) they appeal to us because they evoke our intuitions about what it is correct to say about self-deception, yet (b) they allude to an explanation of self-deception which they fail to spell out, so they need to be supplemented by other elements.

A.O. Rorty, in various articles (e.g. [1972], [1980]) suggests that the unity of the self has been over-emphasised and that selves are loose federations of various elements. Rorty's

work, in particular, contains some allusions to a role element in the description of self-deception.

6. Fox [1973] and [1976] provides a hybrid of Negligence and Radical Interpretation theories, an interesting example of the same theme without a "dissociation" slant. Fox argues that,

> what is meant by "unconscious emotion" is an affect which is both experienced and "disguised" (i.e. self-deceivingly misrepresented by P to himself). (Fox [1973], p413)

Consequently person P can, for example, be afraid even though he does not feel afraid. Fox adds,

> when he talks of repression as a clinically observed phenomenon, Freud is referring to "failed" repression (primarily), which allows affects into consciousness. When it is appropriate to label these as emotions, I should want to say that they appear to P's conscious awareness in a (deliberately) misrepresented form. (p414).

Fox is not arguing that P dissociates his beliefs from his emotions: P experiences the emotion but misrepresents it. I count this as an instance of a "Radical Interpretation" theory of self-deception; the reference to Freud (famous for his interpretative methods) fits in with this view. However the

Freudian theme brings with it a dissociation element. For Freud argues that it is not enough for a psychoanalyst to present and have a patient accept his analysis in order to achieve a "cure" of, for example, a phobia. "Abreaction" must also occur, and abreaction seems to consist in the interpretation offered by the psychoanalyst becoming linked to the affect so that the patient feels emotion (e.g., fear) and does not just hold the theory that the affect is correctly described as fear.

One outstanding debt which I should acknowledge is to Cook [1987]. This may be surprising since Cook's discussion is about "deciding to believe without self-deception" (his title, my emphasis). For me a crucial part of Cook's article was the following passage:

> He set out to engage in a pattern of action, the predictable result of which was his coming to believe a certain theory in biology. In time a hexis developed, and he found himself believing the theory. He learned to believe the theory, where that is not sharply distinguishable from learning that the theory is believable. One might judge that Nick was deceiving himself, but a more charitable account would hold that he has successfully carried out a program of belief acquisition by a roundabout route. (p446)

The expression, "roundabout route" is drawn from Williams [1973], which Cook quotes with approval. I have, perhaps unfairly, used Williams as the butt of many of my criticisms. The tenor of those criticisms, you may recall, was to ask if there is a more direct route by which to acquire beliefs; and my answer to that rhetorical question was that there is no more direct route. So, given my way of reading Cook, it follows that Cook is referring to the process (or one of the processes) by which beliefs in general (not just the self-deceptive ones) are acquired. The process which results in justified true beliefs could, in other applications, result in unjustified false beliefs, whence my claim that there are no special techniques for self-deception.

Cook's article lacks any explanation of how Nick, the exemplar of "deciding to believe without self-deception" achieved his new belief. Nick plans the feat, and then,

> Six years later, Nick emerges from graduate school, Doctorate in hand, fully versed in and accepting of contemporary biological theory. (p443)

How did he do it? Cook does not tell us. He only describes the aftermath:

> Referring to him prior to implementation of his strategy as Nick1, and upon graduation ... as Nick2, this [situation] can be described as follows: Nick1 views his own beliefs as based upon warranting evidence and foresees Nick2's beliefs to be the results of nonevidential causal influences (professors' prestige, social pressures), prejudicially selective habits of attention, and employment of tendentious interpretive hypotheses. Likewise, Nick2 views his own beliefs as evidentially warranted and those of Nick1 as the result of ignorance, distorted perspective, ritual belief reinforcement and other nonrational causal influences. The fact that each can "explain away" the beliefs and the evidential standards of the other provides sufficient "epistemic insulation" between the two to permit Nick2 sincerely to believe p without having to forget or deceive himself about the fact that his belief is a consequence of Nick1's decision to believe that p. (p444)

This is excellent, but does not explain how the trick was done. It strongly implies, though, that it was done by some process (which took six years to perform, incidentally), namely, (a) adopting a role, and (b) interpreting. This invites a Process theory about self-deception, which ties together Role theories and Radical Interpretation theories.

C o n c l u s i o n

I have given a non-paradoxical account of self-deception. To do so, I have rejected some major epistemic theories (such as foundationalism) because, in my view, they give rise to the paradoxes of self-deception and prevent us giving a workable account of what self-deception involves. This is not a trivial result to obtain from a discussion of a "fringe" topic such as self-deception.

My distinction between "data" and "evidence" enables us to simplify considerably our epistemological discussions of evidence and also opens the way to a long overdue reassertion of the importance of pretending (in the archaic sense of that word) for the formation of beliefs and for enquiry generally.

The Process theory of self-deception is the offshoot of a process theory of enquiry. For this reason it enables us to put self-deception in a broader context than it is usual to provide when discussing this topic. Unlike "static" theories of self-deception, the Process theory allows us to give a comprehensive classification of the varieties of self-deception. By listing the "factors of production" it gives clear guidelines for applying the expressions, 'self-deceived'

and 'self-deceiving'. It enables us to contrast and compare self-deception with other related phenomena.

Expressions like 'self-deceived' are part of our language because they have useful work to perform. The Process theory enables us to put them to work more effectively. It also enables us to relate the theories of enquirers as diverse as those in the following list: Plato, Nietzsche, Butler, Marx, Descartes. Sartre, Locke, and Pascal.

Process theories resolve the "must know" problems which give rise to the paradoxes of self-deception. They do not characterise self-deception as a way of having two mutually inconsistent beliefs simultaneously. A self-deceiver may have two (or more) mutually inconsistent interpretations simultaneously; but, at most, only one of these is fixed as a belief at any one time. The self-deceiver may have two or more belief-like interpretations simultaneously, for they may be fixed in different ways e.g. one interpretation may guide actions while another interpretation guides emotions. When someone has two belief-like interpretations, they cannot be fixed in the same respect at the same time; for example, if someone performs an action because he is using interpretation A, he cannot simultaneously not perform the action using interpretation B. Inconsistency within a set of

interpretations is not manifested in inconsistent (i.e. impossible) activity.

Someone who is self-deceived is responsible (in a rather strong sense of 'responsible') for having false beliefs, or not having true beliefs, but most importantly, even if he gains true beliefs and avoids false beliefs, he not only lacks regard for truth, he also displays active preferences for some varieties of falsehoods. Process theories explain how it is possible to be responsible (in this strong sense) for the interpretations one produces and the roles one adopts. Therefore they explain how one can be responsible for one's beliefs or lack of beliefs.

Bibliography

1. Works Cited In The Text

Aristotle. 1972. De Memoria et Reminiscentia, transl. Sorabji, R., in Sorabji, R., Aristotle On Memory (Duckworth)

Audi, R. 1982. Self-Deception, Action And Will, Erkenntnis 18, 159 - 164

Bach, K. 1981. An Analysis of Self-Deception, Phil Phenom Res 41, 351 - 370

Broad, C.D. 1923. The Mind and Its Place in Nature (London: Routledge and Kegan Paul), p363-9

Butler, Upon Self-Deceit, Sermon X, Works (Gladstone edition) Vol II, 142 - 155

Caesar, J. 1951. De Bello Gallico / The Conquest Of Gaul, trans. Handford S.A. (Harmondsworth, Penguin)

Carnap, R. 1962. Logical Foundations of Probability, second edition (Chicago, Illinois, USA: University of Chicago Press)

Cicero. 1948. De Oratore, trans. Rackham, H. (London, Cambridge [Mass], Heinemann, Harvard University Press)

Cook, J.T. 1987. Deciding To Believe Without Self-Deception. Journal Of Philosophy Vol LXXIV No 8, 441 - 446

Demos, R. 1960. Lying To Oneself. J Phil 57

Descartes, R. 1968. Discourse On Method and The Meditations, trans.Sutcliffe, F.E. (Penguin)

Descartes, R. 1970. Philosophical Writings, trans. Anscombe, E. and Geach, P.T. (Nelson's University Paperbacks)

Elster, J. 1983. Sour Grapes (Cambridge, England: Cambridge University Press)

Elster, J. 1979. Ulysses and the Sirens (New York: Cambridge)

Factor, R.L. 1977. Self-Deception And The Functionalist Theory Of Mental Processes, Personalist 58, 115 - 123

Fingarette, H. 1969. Self-Deception (Atlantic Highlands, Humanities Pr)

Fox, M. 1973. On Unconscious Emotions, Philosophy and Phenomenological Research XXXIV, 151 - 170

Fox, M. 1976. Unconscious Emotions: A Reply To Professor Mullane's "Unconscious And Disguised Emotions", Phil Phenomenol Res 36, 412 - 414

Gleich, J. 1988. Chaos: Making A New Science (Heinemann)

Gosse, E. 1964. Father and Son, A Study of Two Temperaments (London, Heinemann Educational Books)

Griffiths, A.P. (ed). 1967. Knowledge and Belief (Oxford University Pr)

Haight, M.R. 1980. A Study Of Self-Deception (Sussex Harvester Pr)

Hamlyn, D.W. 1971 (1). Self-Deception. Aris Soc 45, 45 - 60

Hamlyn, D.W. 1971 (2). Unconscious Intentions, Philosophy 46, 12 -22

Hampshire, s. 1971. "Freedom of Mind", in Freedom of Mind and Other Essays (Princeton, N.J., University Press)

Hellman, N. 1983. Bach On Self-Deception, Phil Phenom Res 44, 113-120

Kipp, D. 1980. On Self-Deception, Phil Quart 30, 305 - 317

Koestler, A. 1958. The Sleepwalkers (Pelican Books 1968)

Locke, J. 1964. An Essay Concerning Human Understanding , abridgement of fifth edition (Fontana)

MacIntyre, A. 1967. A Short History Of Ethics, (London, Routledge)

Martin, M.W. 1979a. Self-Deception, Self-Pretence, And Emotional Attachment, Mind 88, 441 - 446

Marx, K. 1977. Karl Marx Selected Writings, ed. McLellan, D. (Oxford University Press)

Mele, A.R. 1982. 'Self-Deception, Action And Will': Comments, Erkenntnis 18, 159 - 164

Mele, A.R. 1983. Self-Deception, Phil Quart 33, 366-377

Mitchell, W.T.J. 1987. Iconology: Image, Text, Ideology (paperback edition, Chicago, University of Chicago Pr)

Morris, P.S. 1980. Self-Deception: Sartre's Resolution Of The Paradox, in Silverman, H.J. (Ed), Jean-Paul Sartre, 30 - 49

Mounce, H.O. 1971. Self-Deception, Aris Soc 45, 61 - 72

Nietzsche, F. 1964. The Complete Works Of Friedrich Nietzsche, ed. Oscar Levy (New York)

Nietzsche, F. 1968. Will To Power, trans. Kaufmann, W. (Vintage Books)

Nietzsche, F. 1969. On The Genealogy Of Morals and Ecce Homo, trans. Kaufmann, W. (Vintage Books)

Nietzsche, F. 1973. Beyond Good and Evil, Prelude To A Philosophy of the Future, trans. Hollingdale, R.J. (Penguin)

Nietzsche, F. 1974. The Gay Science, trans. Kaufmann, W. (Vintage Books)

Onions, C.T. (ed). 1983. The Shorter Oxford English Dictionary (book Club Associates by arrangement with Oxford University Press)

Pascal, B. 1966. Pensees, trans. Krailsheimer, A.J. (Penguin)

Pears, D. 1974. The Paradoxes Of Self-Deception Teorema Mono 1, 7 - 24

Pears, D. 1984. Motivated Irrationality (Oxford Clarendon Pr)

Penelhum, T. 1964. Pleasure and Falsity, American Philosophical Quarterly 1, 81 - 91

Peterman, J. 1983. Self-Deception And The Problem Of Avoidance, S J Phil 21, 565 - 574

Plato. 1949. Theaetetus, transl. Jowett, B. (Bobbs-Merill)

Plato. 1974. The Republic ,transl. Lee, D. (Penguin)

Pugmire, D. 1969. 'Strong' Self-Deception, Inquiry 12, 339 - 361

Rorty, A. O. 1972. Belief And Self-Deception, Inquiry 15, 387 - 410

Rorty, A.O. 1980. "Akrasia" And Conflict, Inquiry 23, 193 - 212

Russell, J.M. 1981. Reflection And Self-Deception, Res Phenomenol 11, 62 - 74

Santoni, R.E. 1978. Bad Faith And 'Lying To Oneself' Phil Phenomenol Res 38, 384 - 398

Sartre, J.P. 1947. No Exit, trans. Gilbert, S. (New York, Knopf)

Sartre, J.P. 1975. Being and Nothingness, trans. Barnes H.E. (New York: Washington Square Pr)

Silver, M., Sabini, J. and Miceli, M. 1989. On Knowing Self-Deception, Journal for the Theory of Social Behaviour 19:2

Solomon, R.C. 1976. The Passions (Garden City NY Anchor Pr)

Solomon, R.C. 1978. Phony Feelings, Journal of Philosophy 78, 697 - 699

de Sousa, R.B. 1970. Review of Self-Deception, Inquiry 308 - 321

de Sousa, R.B. 1978. Self-Deceptive Emotions, Journal of Philosophy 75, 684 - 697

Sperry, R.W. 1961.Cerebral Organisation and Behaviour, Science Vol 133, 1749 - 1757

Von Oech, R. 1990. A Whack On The Side Of The Head (UK edition, Thorsons)

Whisner, W. 1989. Self-Deception, Human Emotion, and Moral Responsibility: Towards A Pluralistic Conceptual Scheme, Journal For The Theory Of Social Behaviour 19:4

Wilkes, K.V. 1978. Consciousness And Commissurotomy, Philosophy 53, 185 - 199

Williams, B. 1973. "Deciding To Believe", in Problems of the Self (New York: Cambridge)

Winters, B. 1979. Believing at Will, Journal of Philosophy 79, 243- 256

2. Other Philosophical Works On Self-Deception

Abelson, R. 1977. Persons: A Study In Philosophical Psychology (London. Macmillan)

Alexander, P. 1974. Wishes, Symptoms And Actions, Aris Soc 48, 119 - 134

Audi, R. 1975. The Epistemic Authority Of The First Person, Personalist 56, 5 - 15

Audi, R. 1976. Epistemic Disavowals And Self-Deception, Personalist 57, 378 - 385

Barnes, H.E. 1959. The Literature Of Possibility: A Study In Humanistic Existentialism (Lincoln Univ Of Nebraska Pr)

Black, M. 1982. The Prevalence Of Humbug, Phil Exch 3, 1 - 24

Boyers, R. 1974. Observations On Lying And Liars, Rev Exist Psych Psychiat 13, 150 - 168

Burks, D.M. 1970. Persuasion, Self-Persuasion And Rhetorical Discourse, Phil Rhet 3, 109 - 119

Canfield, J.; McNally, P. 1961. Paradoxes Of Self-Deception, Analysis 21, 140 - 144

Canfield, J.V.; Gustavson, D.F. 1962. Self-Deception, Analysis 23, 32 - 36

Carrier, J.G. 1979. Misrecognition And Knowledge. Inquiry 22, 321 - 342

Champlin, T.S. 1976. Double Deception, Mind 85, 100 - 102

Champlin, T.S. 1977. Self-Deception: A Reflexive Dilemma, Philosophy 52, 281 - 299

Champlin, T.S. 1979. Self-Deception: A Problem About Autobiography, Aris Soc 53, 77 - 94

Champlin, T.S. 1984. Self-Deception In Second-Rate English, Philosophy 59, 259-261

Cioffi, F. 1974. Wishes, Symptoms And Actions, Aris Soc 48, 97 - 118

Cohen, A. 1982. Kierkegaard As A Psychologist Of Philosophy, J Brit Soc Phenomenal 13, 103 - 119

Cosentino, D.A. 1980. Self-Deception Without Paradox, Phil Res Archive 6, No 1388

Cronin, R.G. 1977. A Definition Of Believing, Auslegung 4, 122 - 132

Daniels, C.B. 1974. Self-Deception And Interpersonal Deception, Personalist 55, 244 - 252

Davis, W.H. 1971. The Freewill Question (The Hague. Nijhoff)

Deutsch, E. 1982. Personhood, Creativity And Freedom (Honolulu Univ Of Hawaii Pr

Dilham, I., and Phillips, D.Z. 1971. Sense And Illusion (NY Humanities Pr)

Doore, G.L. 1983. William James And The Ethics Of Belief, Philosophy 58, 353 - 364

Erickson, S.A. 1984. Human Presence: At The Boundaries Of Meaning (Macon Mercer Univ Pr)

Evans, D. 1975. Moral Weakness, Philosophy 50, 295 - 310

Exdell,J.;Hamilton,J. 1975. The Incorrigibility Of First Person Avowals, Personalist 56, 389 - 394

Festinger, L. 1957. A Theory of Cognitive Dissonance (Stanford, California: Stanford University Press)

Flanagan, K. 1981. The Experience Of Innocence As A Social Construction, Phil Stud (Ireland) 28, 104 - 139

Flew, A. 1980. Parapsychology: Science Or PSeudo-Science? Pac Phil Quart 61, 100 - 114

Foss, J. 1980. Rethinking Self-Deception, Amer Phil Quart 17, 237 - 242

Frazier, A.M. 1977. F.H. Bradley's Analysis Of Religious Consciousness, Ideal Stud 7, 239 - 251

Fuller, G. 1976. Other-Deception, SW J Phil 7, 21 - 31

Gardiner, P.L. 1970. Error Faith And Self-Deception, Proc Aris Soc 70, 197 - 220

Govier, T. 1983. Nuclear Illusion And Individual Obligations, Can J Phil 13, 471 - 492

Guthrie J.L. 1981. Self-Deception And Emotional Response To Fiction, Brit J Aes 21, 65 - 74

Hare, R.M. 1952. The Language Of Morals (Oxford Clarendon Pr)

Harris, J. 1981. Ethical Problems In The Management Of Some Severely Handicapped Children, J Med Ethics 9, 117 - 119

Hausman, C.R. 1966. The Existence Of Novelty, Phil Forum (Pacific) 4, 3 - 60

Hausman, C.R. 1967. Creativity And Self-Deception, J Existent 7, 295 - 308

Heil, J. 1983. Believing What One Ought, J Phil 80, 752 - 764

Kellenberger, J. 1970. Religious Discovery, Sophia 9, 22 - 33

Kellenberger, J. 1972. Religious Discovery, Faith, And Knowledge (Englewood Cliffs NJ Prentice-Hall)

Kellenberger, J. 1980. The Death Of God And The Death Of Persons, Relig Stud 16, 263 - 282

Ketchum, S.A. 1981. "Moral Redescription And Political Self-Deception" in Vetterling-Braggin, M. (Ed) Sexist Language, 279 - 289, (Totowa, Littlefield Adams)

King-Farlow, J. 1963. Self-Deceivers and Sartrean Seducers, Analysis 131 - 136

King-Farlow, J. 1973a. Critical Notice of Herbert Fingarette's Self-Deception, Metaphilosophy 4, No.1, 76 - 84

King-Farlow, J. 1973b. Bonne Foi, Mauvais Foi, Sincerite et Espoir, Dialogue 12, No. 3, 502 - 514

King-Farlow, J. 1978. Philosophical Nationalism: Self-Deception And Self-Direction, Dialogue 17, 591 - 615

King-Farlow, J. 1981. Self-Mastery And The Master Self, Pac Phil Quart 62, 47 - 60

King-Farlow, J. 1981. Deceptions, Assertions, Or Second-String Verbiage?, Philosophy 56, 100- 105

Kittay, E.F. 1982. On Hypocrisy, Metaphilosophy 13, 277 - 289

Kovar, L. 1974. The Pursuit Of Self-Deception, Rev Exist Psych Psychiat 13, 136 - 149

Linehan, E.A. 1982. Ignorance, Self-Deception, And Moral Accountability, J Val Inq 16, 101 - 116

Marko, Kurt. 1974. Some Remarks On Expectations Of Imminent Changes In Socialist Countries, Stud Soviet Tho 14, 257 - 261

Martin, M.W. (ed) 1986. Self-Deception and Self-Understanding (Lawrence, Kansas, University of Kansas Press)

Martin, M.W. 1977. Immorality And Self-Deception: A Reply To Bela Szabados' "The Morality Of Self-Deception", Dialogue (Canada) 16, 245 - 273

Martin, M.W. 1978. Sartre On Lying To Oneself, Phil Res Arch 4, No 1252

Martin, M.W. 1979b. Factor's Functionalist Account Of Self-Deception, Personalist 60, 336 - 342

Martin, M.W. 1979c. Morality And Self-Deception: Paradox, Ambiguity, Or Vagueness?, Man World 12, 47 - 60

Martin, M.W. 1984. Demystifying Doublethink: Self-Deception, Truth and Freedom in "1984", Soc Theor Pract 10, 319-331

McLaughlin, B., and Rorty, A. (eds). 1988. Perspectives On Self-Deception (Berkeley, California: University Of California Pr)

Mele, A.R. 1987. Recent Work On Self-Deception. American Psychological Quarterly, 24 (I) 1 - 17

Miri, M. 1974. Self-Deception, Philosophy and Phenomenological Research 34, p577

Molina,F. 1962. Existentialism As Philosophy (Englewood Cliffs NJ Prentice-Hall)

Mullane, H. 1976. Unconscious and Disguised Emotions, Philosophy and Phenomenological Research

Mullen, J.D. 1981. Kierkegaard's Philosophy: Self-Deception And Cowardice In The Present Age (NY Mentor Book)

Oates, J.C. 1974. The Imposters, Rev Exist Psych Psychiat 13, 169 - 183

Palmer, A. 1979(1). Self-Deception: A Problem About Autobiography, Aris Soc 53, 61 - 76

Palmer, A. 1979 (2). Characterising Self-Deception, Mind 88, 45 - 58

Paluch, S. 1967. Self-Deception, Inquiry 10, 268 - 278

Paskow, A. 1979. Towards A Theory Of Self-Deception, Man World 12, 178 - 191

Paton, H.J. An Alleged Right To Lie: A Problem In Kantian Ethics, Kantstudien 45, 190 - 203

Paul, R. 1982. Teaching Critical Thinking In The "Strong" Sense: A Focus On Self-Deception, World Views, And A Dialectical Mode Of Analysis, Inform Log 4,2 - 7

Pears, D.F. 1975. Questions In The Philosophy Of Mind (London. Duckworth)

Pole, D. 1971. The Socratic Injunction, J Brit Phenomenol 2, 31 - 40

Pole, D. 1974. Virtue And Reason, Aris Soc 48, 43 - 62

Quinton, A. 1974. Sobre La Definicion Del Conocimiento, Teorema 4, 159 - 175 (note: Spanish language)

Radden, J. 1984. Defining Self-Deception, Dialogue (Canada) 23, 103-120

Reilly, R. 1976. Self-Deception: Resolving The Epistemological Paradox, Personalist 57, 391 - 394

Roberts, P. 1975. The Psychology Of Tragic Drama, Ideas And Forms In English Literature, John Lawler (Ed), (Boston Routledge Kegan Paul)

Russell, J.M. 1978. Saying, Feeling, And Self-Deception, Behaviourism 6, 27 - 43

Saunders, J.T. 1975. The Paradox Of Self-Deception, Phil Phenomenol Res 35, 559 - 570

Scott-Taggart, M.J. 1972. Socratic Irony And Self Deceit, Ratio 14, 1 - 15

Shapiro, G. 1974. Choice And Universality In Sartre's Ethics, Man World 7, 20 - 36

Sharpe, R.A. 1975. Seven Reasons Why Amusement Is An Emotion, J Value Inq 9, 201 - 203

Siegler, F.A. 1968. An Analysis Of Self-Deception, Nous 2, 147 - 164

Siegler, F.A. 1968. An Analysis Of Self-Deception, Nous 2, 147 - 164

Stack, G.J. 1983. The Sartrean Self: Ambivalent Or Paradoxical? Phil Exch 14, 121-127

Szabados, B. 1973. Wishful Thinking and Self-Deception, Analysis 33, p205

Szabados, B. 1974a. Rorty On Belief And Self-Deception, Inquiry 17, 464 - 473

Szabados, B. 1974b. Self-Deception, Can J Phil 4, 41 - 49

Szabados, B. 1974c. The Morality of Self-Deception, Dialogue 25 - 34

Szabados, B. 1977. Fingarette On Self-Deception, Phil Papers 6, 21 - 30

Szabados, B. 1979. Hypocrisy, Can J Phil 9, 195 - 210

Siegler, F.A. 1962. Demos On Lying To Oneself, Journal of Philosophy 54, p474

Siegler, F.A. Self-Deception, Austl J Phil 41, 29 - 43

Thielst, P. 1976. Poul Martin Moller (1794 - 1838): Scattered Thoughts, Analysis Of Affectation, Combat With Nihilism, Dan Yrbk Phil 13, 66 - 83

Watson, G. 1978. Appropriate Emotions, Journal of Philosophy 78, 699

Wilshire, B. 1972. Self, Body And Self-Deception, Man World 5, 422 - 447

Wilson, K. 1980. Self-Deception And Psychological Realism, Phil Invest 3, 47 - 60

Wollheim, R. 1971. Freud (London. Collins)

Young, R. 1980. Autonomy And The 'Inner Self', Amer Phil Quart 17, 35 - 43

Zimmerman, B.K. 1977. Self-Discipline Or Self-Deception?, Values Ethics Health Care 2, 120 - 132