1

CANADIAN THESES ON MICROFICHE

I.S.B.N.

THESES CANADIENNES SUR MICROFICHE

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

93

0-315-12064-9

National Library
of CanadaBibliothèque nationale
du Canada

Canadian Theses Division Division des thèses canadiennes

Ottawa, Canada
K1A 0N4

59977

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

• Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

IAN BRUCE CRAIB

Date of Birth — Date de naissance

DEC 13, 1955

Country of Birth — Lieu de naissance

CANADA

Permanent Address — Résidence fixe

58 JEFFERSON AVE.
PTE. CLAIRE, QUÉBEC

Title of Thesis — Titre de la thèse

SELF-DECEPTION

University — Université

CARLETON

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

M. A.

Year this degree conferred — Année d'obtention de ce grade

1982

Name of Supervisor — Nom du directeur de thèse

STANLEY G. CLARKE

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication, ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date

Oct. 18, 1982

Signature

I. B. Craib

SELF-DECEPTION

by

© Ian B. Craib

A thesis submitted to the Faculty of
Graduate Studies and Research in
partial fulfillment of the requirements
for the degree of Master of Arts.

Department of Philosophy
Carleton University
Ottawa, Ontario
September 1982

The undersigned recommend to the Faculty of
Graduate Studies and Research acceptance of
the thesis

"Self-Deception" submitted by Ian B. Craib
in partial fulfillment of the requirements
for the degree of Master of Arts.

.....
Thesis Supervisor

.....
Chairman
Department of Philosophy

ABSTRACT

The concept of self-deception has been thought to embody a paradox of knowing ignorance. I argue that whether or not the concept is inconsistent, we need an explanation of the causal structure of the phenomena of self-deception if we wish to ascertain whether it makes sense to hold the self-deceiver responsible for his deceived state. Traditional accounts have had difficulty reconciling the intentionality of deception, which is supposedly required if attribution of responsibility is to be appropriate, with the innocence of one who is deceived. The concept of attentional focus is often employed to this end, but it is shown to be a crude and inadequate explanation. I develop an explanation of self-deception as a dissociation of memories incompatible with a powerful emotion, citing empirical evidence of the mood-state-dependency of recall. The self-deceiver does not avoid the evidence contrary to his cherished belief: the evidence simply fails to transfer to an incompatible mood-state. I argue further, that self-deception can be avoided if one adheres to sound epistemic policies of belief management, and that consequently, falling into self-deception is culpable cognitive negligence.

Thanks to Stan Clarke

and

Barbara Abouziane

and

Helen Smith
(to whom I have been of
very little assistance of late)

TABLE OF CONTENTS

CHAPTER 1: STRATEGY AND CLASSIFICATION	1
Strategy	1
Types of Philosophical Accounts	11
Survey	13
Notes to Chapter 1	21
CHAPTER 2: MOTIVATED ERROR THEORIES.	22
Canfield and Gustafson	22
Sartre	23
Poff-Cliffton	29
Notes to Chapter 2	36
CHAPTER 3: INTENTIONAL DECEIVER THEORIES AND THE CONCEPT OF ATTENTIONAL FOCUS	38
Fingarette on Mechanism	41
Rorty on Mechanism	65
Notes to Chapter 3	77
CHAPTER 4: MOOD-DISSOCIATION AND SELF-DECEPTION	81
Mood-State-Dependency Effects	82
Consequences for Memory Theory	98
Application to Self-Deception	109
Notes to Chapter 4	116
CHAPTER 5: SELF-DECEPTION AND RESPONSIBILITY	120
Notes to Chapter 5	131
BIBLIOGRAPHY	132

CHAPTER I

STRATEGY AND CLASSIFICATION

Strategy

Let us suppose that Leonard, a nationalistic American of my acquaintance, believes that Ronald Reagan is a man dedicated to world peace. If I believe that Leonard is deceiving himself, I will probably express it like this: "Leonard, though he is quite well-informed and ordinarily very rational, and in spite of all the evidence to the contrary, appears to believe that Reagan is a peace-maker. He must be deceiving himself". 'Self-deception' functions here as an explanatory term: it is not merely a description of Leonard's cognitive state, but an account of why he is in that particular sort of state. The unjustified belief, in the presence of sufficient counter-evidence and an apparent capacity to draw the correct inference, is a puzzling condition which we explain by appealing to the common notion of self-deception.

Further support for the contention that 'self-deception' is explanatory is obtained by examining its use more closely. If an account of one thing is to function as an explanation of something else, it must add something to the account of the something else which is not present in a mere description of the latter. In addition, an explanation is most commonly advanced in the interest of a practical purpose for which a mere description would not suffice. Both of these conditions

appear to be satisfied in the use of 'self-deception'. The common notion of self-deception is in fact a crude sort of theoretical account of what causes the kinds of cognitive states which induce ascriptions of 'self-deception': it alleges that self-deceivers hold unjustified beliefs because they wish to believe them, and that the means by which this is accomplished has some similarity to the means we use to deceive others. The problems with this crude theory are well known,¹ but it is nonetheless a theory, and adds content beyond a mere description of the cognitive states concerned. Explanations by self-deception serve the practical purpose of grounding attribution to self-deceivers of some sort of responsibility for their erroneous beliefs. These attributions are often problematical with regard to their propriety and vague in extent, largely because of the difficulties with the theoretical content of the common notion of self-deception. But we do commonly wish to attribute some sort of responsibility, and it is not obvious that it would be possible even in principle to do this merely on the basis of description of the cognitive states involved, since attributions of responsibility normally require evidence of agency and any suggestion of which would have to be inferred as an explanation of observed behaviour, and therefore not purely descriptive.²

We shall take it, then, that 'self-deception' is an explanatory term. We shall next ask: what is the meaning

of the term, to what does it refer? A detailed discussion of theories of meaning would be out of place, and is beyond the scope of an investigation of this sort. I shall instead adopt what I take to be a familiar option with only a brief discussion of why I think this stance appropriate to the case of self-deception.

We have seen that the common notion of self-deception includes a quasi-theoretical explanation, and the survey of philosophical accounts which is to follow will introduce other, more detailed forms of explanation. If these theorists are to be taken to be giving accounts of the same phenomena as are users of common language, it would seem reasonable to conclude that 'self-deception' refers to the actual deep structure of these phenomena. An objective of a philosophical account will then be to provide a more accurate picture of this structure than is present in the common notion. We shall be realists: 'self-deception' will be taken to refer to the causal structure which explains certain kinds of cognitive states. Of course, an account is owed of just what kinds of cognitive states we are talking about here: we shall have to outline the defining features of self-deception, which identify what the causal structure is a structure of. But the reference of 'self-deception' reaches beyond these, just as the reference of 'measles' reaches beyond the symptomatology of spots, nausea, and high temperature, to the activities of the bacteria which

cause the symptoms. The defining features of self-deception may be seen as the symptoms of the actual referent. I am not a medical scientist, and so I don't know very much about the actual causal structure of measles. But this knowledge is a part of our shared culture, and it is to this causal structure that my use of 'measles' refers, even though the defining features of measles for me and for the majority of language-users are a fairly simplistic version of its symptoms.³ In the case of 'self-deception' less is known about the cause or causes, but we shall still maintain that it refers to the as yet unknown causal structure of the kinds of cognitive states that it is employed to explain.

What are these "certain kinds of cognitive states"? Basically those in which:

- (i) Either: (a) a belief appears to be held, in spite of fairly conclusive evidence against it, of which the affected person could be expected to be aware; or (b) a belief appears not to be held, in spite of fairly conclusive evidence in its favour, of which the affected person could be expected to be aware.
- (ii) The belief avoided, or maintained, is one which the affected person might be expected to want to avoid, or maintain.
- (iii) The belief, and the evidence it flouts, concern an issue associated with states of emotional arousal in the affected person.

The above are not intended as necessary or sufficient conditions for the applicability of the common notion of self-deception, but as provisionally definitive features of the concept of self-deception as we use it, which are defeasible, however, by a theoretical account of the nature of the referent of 'self-deception'. They are meant to be as widely acceptable as possible, if not exactly theory-neutral.

When we say of someone that he is fooling himself, we do not necessarily mean that he "knows" of the existence and strength of the evidence against his belief: it is rather, as Siegler argues, that we think that he ought to know this.

Consider the following: It is widely known that Brown's wife has been unfaithful to him. It is perfectly obvious to everybody -- but Brown. Brown tells his friend that he is beginning to get worried about the relationship between his wife and her friend. He says that unless she breaks off that relationship she may be led to acts of infidelity. White is puzzled at what he considers either naivete or stubborn pride in Brown. "Look, Brown" he says, "you are deceiving yourself if you think that your wife has been faithful to you all this time".⁴

Siegler argues that in this case White suggests not that Brown believes what he knows to be false, but that he holds an erroneous belief which it is unreasonable to have, and that White's point is to urge Brown to reconsider the evidence.⁵

It is not that the "deceived" knows that his belief is (or was) unwarranted; the point indicated is that he ought to know (have known) better.⁶

Ascriptions of self-deception are characteristically made in cases where the affected person appears to hold an unwarranted belief, or appears not to hold an inescapable belief, but fall short of requiring that he actually knows that this is the case.

The second feature, that of apparent motive, is virtually unchallenged in the literature -- even Penelhum admits that "motives are, however, usual"? It would seem unobjectionable and unavoidable to say that there is at least a temptation to ascribe a motive regarding the cognitive states for which self-deception is offered as an explanation.

The third feature cited is implicit in the literature, if not always explicitly stated. 'Self-deception' is used in cases where the beliefs concerned are ones about which the affected person might be expected to care a lot -- witness the example immediately above. This does not reduce to motive: one may imagine motives for believing that it will be sunny tomorrow in spite of the forecast, but it would be difficult to imagine self-deception regarding it, unless by some chance there were emotionally significant issues dependent on its being sunny. The classic cases of self-deception involve love, self-worth, or self-identity.

'Self-deception' is, then, an explanatory term which refers to the cause or causes of the kinds of cognitive states outlined above. The requirements of a philosophical

account of self-deception which follow from what has been said so far are as follows: first, we will need a better explanation of the defining features of self-deception than is provided in the common notion; and second, we will require an explication of whatever consequences that this improved explanation will have for proper practical use of the term. The problems with the common notion stem from a conflict within that model between the two poles of the self-deceiver's-dual status as "deceiver" and "deceived". There is a certain implausibility to this as an account of a cognitive state: as the deceiver, bearer of the want to believe in a certain way and presumably of the intention to cause such a belief, the self-deceiver would appear to be necessarily aware of at least his motivated reasons for holding the belief, if not actually of the fact that it is unwarranted. But as the one who is deceived, it would seem that the self-deceiver could not be aware of any of this without abandoning the ranks of the "deceived". Whether or not the self-deceiver can plausibly be considered to be the agent of deception is indeterminate in the common notion, and this is a problem both from the standpoint of theory and in terms of the practical usefulness of the concept in grounding attributions of responsibility.

Our approach in giving a philosophical account of self-deception will be to produce a theory of the nature of the referent of 'self-deception' and examine its implications for

practical use of the concept. The first step is to identify the kinds of cognitive states that it is used to explain, and we have done this. The next is to develop a psychological theory to explain the causal structure of those states, and this will be attempted in the fourth chapter. The final step, explication of the proper practical use of the concept in view of the nature of the referent, will be the business of the final chapter. A preliminary question, however, which is relevant to both our own approach and to the survey of the philosophical literature on self-deception which is shortly to follow, concerns the acceptable bounds of revision of the defining features of self-deception and of the field of cases to which it applies.

How much, and what sort of revision of the defining features and of the field of cases is permissible in a philosophical account of self-deception? The short answer is that revisions demanded by a theory of the nature of the referent are justified, whereas those based on purely conceptual grounds generally are not.

Faced with defining features of common use which are indeterminate regarding the agency of the self-deceiver, and with a field of cases of self-deception in which some cases appear more deliberate than others, some conceptual analysts have attempted to make the concept more precise by stipulating that only those cases which literally fit the deliberate deception model ought properly to be called

'self-deception'. The advantage of this, of course, is that the concept is thereby made more precise, and its application more consistent. Unfortunately, clarity and consistency are insufficient grounds for the excisions made: there is no independent reason to favour cases which appear to be more deliberate, for it is not clear that they are in any sense more central than the cases which fit more naturally into the model of the "deceived victim". The indeterminacy of the defining features of self-deception cannot be resolved by definitional fiat.

A perceived need to distinguish self-deception from related concepts such as wishful thinking or simple ignorance is often advanced as another ground for revising or delimiting the defining features and field of cases of self-deception, once again in order to make the concept more precise. But this, too, is unjustified without further reason, for it is surely necessary first to determine whether these concepts do in fact refer to different things, and that is a conclusion which can only be justifiably asserted after a thorough investigation of the purported referents. Such an investigation will be largely empirical in the case of 'self-deception'.

Conservative revision may be justifiable where required by an empirical theory of the nature of the referent of 'self-deception'. Large scale revision of the defining features and delimitation of the field of cases would be

highly unlikely, since a demand on an empirical theory is that it be applicable to at least a large majority of the field of cases that it purports to explain, but because the make up of the field of cases and defining features may be dependent to some extent on the quasi-theoretical content of the common notion of self-deception, some adjustments may be required, and are defensible by the normal principles of scientific theory-succession. For example, it might turn out that a certain causal structure is sufficient to explain the large majority of cases of what has been called self-deception, but not all of them. In this case we would be justified in claiming that the exceptions were not really self-deception, but something else. Or we might find that one of the defining features of self-deception was inessential or misleading, and that its preliminary role in defining self-deception might be better accounted for in another way. And if other cognitive states were to turn out to share much the same causal structure, we might wish to admit ~~them~~ to the field of cases of self-deception. These types of ~~revision~~ are in no sense arbitrary: they are nationalized by empirical theory.

We shall make the following requirements of a philosophical account of self-deception.

- (i) It must be applicable to the greater part of the field of cases of self-deception.
- (ii) It must be internally coherent.

- (iii) It should include or be complementary to a plausible theoretical explanation of the referent.
- (iv) It should give an account of the key issue of whether or not the self-deceiver is responsible for his state.

The first two are no more than is asked of any scientific model, and 'plausible' in the third, is to be understood as it would be in the case of a scientific model: as simple, persuasive, and open to refutation. This third requirement is justified by the evident advantage in basing philosophical explication on a scientifically acceptable theoretical explanation of the referent, rather than the vague theory implicit in ordinary language. The fourth requirement stands as the basic question to be answered in a philosophical account of self-deception. The issue of whether or not a self-deceiver is to be taken to be responsible for his state of ignorance is a serious indeterminacy in the common sense view. In my opinion, clarification of this issue is the *raison d'être* of the philosophical problem of self-deception.

Types of Philosophical Accounts

There are two general types of philosophical accounts: those which I shall call *deceiver theories*, which follow to some extent the analogy with deception of others, and *error theories*, which reject that analogy

as misleading and/or inherently paradoxical. Deceiver theories for the most part postulate that the self-deceiver holds two incompatible beliefs, though they need not be of the same type. Error theories reject this, characterising self-deception as a single belief in peculiar circumstances.

The deceiver theory group is further subdivided according to whether or not the deception of the self by the self is taken to be played out at a conscious level. If the scope of what is to be referred to as 'intentional' is restricted to purposeful acts involving some degree of conscious deliberation, we may refer to these alternative types as intentional and non-intentional deceiver theories.

The error theory group can also be profitably subdivided. Two issues determine the subdivisions: the issue of whether or not motive must play an essential part in the explanation of the sorts of error characteristic of self-deception, and that of whether or not an affected person can be construed as responsible for his error in a theory which eschews motive as an essential feature. The three types of error theory that I identify are: motivated error theories, conflict-state theories, and "Ought-to-know" error theories.

It should be noted here that this classification concerns strategies rather than individual accounts. Rorty,⁹ for example, invokes elements of both motivated error theory and intentional deceiver theory in her account of self-deception.

Survey of Philosophical Accounts

Of the five types of philosophical account to be considered, I shall have the least to say about non-intentional deceiver theories. An exhaustive treatment of psychoanalytic theories could not be attempted in a thesis concerned primarily with analytic accounts, and with advancing a different sort of psychological justification for an account of self-deception. I shall, however, offer a thumbnail sketch of explanations in terms of unconscious processes, and offer a couple of general criticisms of this approach.

Non-intentional deceiver theories hypothesize that there is a "faculty" of the mind which serves a censoring function, filtering out hazardous or undesirable information about oneself and the external world, without the conscious self becoming aware of its activities.

An account of this type might be internally coherent, since it appears to avoid the paradox of knowing ignorance by the division of the self into conscious and unconscious parts,¹⁰ and it would also appear that a definite (and negative) answer has been given to the question of the self-deceiver's responsibility for his state. It would seem, however, that an explanation in terms of unconscious processes would be less satisfactory regarding the other two requirements of philosophical accounts previously given.

In the first place, an account which would put the "deceiver" out of range of the conscious self could be expected

to encounter some difficulty in accounting for cases in which self-deception appears to have been overcome by the victim's own efforts. Self-defeasible self-deception would appear to be fairly common: we often say things like "I was just deceiving myself when I thought that I would write something original in philosophy", referring to a time when one actually believed something now regarded as silly and unreasonable. Presumably, one would not require psychoanalysis in order to arrive at such a realization -- a lifetime enrollment in the school of hard knocks would appear to be enough. In non-intentional deceiver theories, however, the censor precedes conscious awareness, and so it is hard to see how a non-intentional self-deceiver (of the unconscious process type under discussion here) could ever overcome the deception by himself. This runs non-intentional deceiver theories afoul of the requirement that a philosophical theory apply to the greater part of the class of cases.

The more serious difficulty, however, is that an account which hypothesizes an unconscious mechanism in the absence of corroborative evidence in its favour, is inferior, in principle, on the grounds of unnecessary complexity, to an account which can do without such purely explanatory structures. Theories appealing to the unconscious violate our third requirement of a philosophical account: that it include or be consistent with a plausible theoretical explanation of the referent, for the psychological theories

backing them stand under the shadow of Occom's Razar, and so cannot be considered to be plausible, so long as there are realistic alternatives which do not need to invoke unconscious processes.

In chapters four and five I shall outline the beginnings of a psychological theory capable of providing just such an alternative.

In saying that non-intentional deceiver theories are inferior in principle, I am not recommending outright rejection. If attempts to account for self-deception without recourse to the unconscious continue to be unsatisfactory, or if persuasive corroborating evidence of the operation of unconscious processes or of the existence of the right sort of brain structures comes to light, there will be good reason to consider them as providing potentially the most fruitful avenues of investigation.

Philosophical accounts of the intentional deceiver type make up the bulk of the analytical literature on self-deception, and as a number of them will be considered in detail in the next chapter, I shall confine my comments here to a brief outline of the structure of this sort of account, and a few words concerning a characteristic strategy for avoiding an inherent difficulty.

Intentional deceiver theories start out with the basic claim that in self-deception the afflicted party has intentionally, in the sense of a purposeful act or series of acts involving some degree of conscious deliberation, brought

it about that he is deceived. There is some uncertainty amongst the philosophers who hold this view as to how much conscious deliberation is involved, but they share a commitment to the idea that self-deceivers are to at least some extent directly responsible for their state of ignorance, and that this responsibility is due to the intentional character of the process which leads to the eventual state. This has seemed the most natural way to elaborate the common use of 'self-deception', but it also leads directly into paradox. It is likely for this reason that in a recent MA thesis on self-deception, Deborah Poff-Clifton characterises this group of models as "paradox-embracing".

'Embracing', however, is not a terribly sympathetic description of what these theorists (i.e. Demos, Pugmire, Fingarette) are trying to do. The paradox is indeed the basic problem with which accounts of this type must come to grips, but most of the philosophers in this group have made substantial efforts to avoid it, by introducing the concept of the focussing of attention.

The idea that self-deception is to be understood in terms of attentional focus is currently taken to be one of the most promising approaches. I will direct the whole of the third chapter to it, since it is the main alternative to the sort of account I intend to advance.

Consideration of the three types of error theories will make up the balance of this chapter. These theories share a rejection of the notion of analysing self-deception analogously with the deception of others, in the sense that there is no attempt to locate a deceiver. Self-deception is treated as a kind of dysfunction of the belief-formation process.

Motivated-error theories are the most common of the three types, and they advocate an analysis of self-deception as a kind of "belief in the face of the evidence", in which there is a perversion of normal belief management strategies by motives. I shall deal with theories of this type in detail in the next chapter, but it would appear that their general formulation breaks down into either the intentional deceiver or unconscious process theory, depending on how the action of the corrupting motive is to be understood.

The second variety of error theory, of which Penelhum appears to be the only adherent, is the conflict-state theory. Penelhum states that his objective, which he believes is the only plausible one, is to provide "a consistent description of inconsistent behaviour".¹² This is different from the objectives of the other theories that we have and will consider, and from the theory which we will advocate, all of which aim at producing an account of self-deception that demonstrates an underlying consistency in apparently inconsistent behaviour.

Penelhum suggests that self-deception can be described in the following way:¹³

- (i) Subject holds belief in the face of very strong evidence against it.
- (ii) Subject has knowledge of the evidence against his belief.
- (iii) Subject has knowledge of the force of the contrary evidence.

The features of this description are not to be thought of as merely identifying different possible ascriptions which might be made of the self-deceiver. It is not merely that he asserts the belief, but that he in some sense holds it; not merely that we might expect him to know of the existence and force of the contrary evidence, but that he does know.¹⁴

Penelhum argues that since features (ii) and (iii) amount to the same thing as belief¹⁵ - knowledge of the evidence and of its force - self-deception is to be construed as a conflict-state in which conflicting views are simultaneously held to be true, with the consequence that we cannot properly ascribe belief at all.¹⁶

This description may be consistent, but it is far from acceptable as a conceptual model, being mysterious psychologically, and apparently more appropriate as a description of madness than of self-deception.

It is, for a start, rather difficult to imagine what it would be like to be in a conflict-state: a state in which both belief and disbelief in a proposition might be ascribed simultaneously. It might seem as if some sense might be made of this with the idea of one's accepting one belief, and not the other, presumably on non-epistemic grounds; but Penelhum specifically rejects this possibility.¹⁷ In a purely epistemic sense, it is hard to see what else "partial belief" could amount to other than perplexity. This would not seem to be what is meant by 'self-deception', as it does not involve any deception or error at all. But in any case, even if it could be shown that the boundary between perplexity and self-deception were less than crystal clear, Penelhum rejects this possibility as well in discussing Canfield and Gustafson.¹⁸ So we are somewhat at a loss here, unless what Penelhum has in mind is a state a good deal closer to outright madness.

Patrick Gardiner suggests that madness is what a mental state fitting Penelhum's model would amount to,¹⁹ and argues that (ii) and (iii) are too strong to accurately describe self-deception. This is fair enough: actual knowledge of the existence and force of contrary evidence is not an essential part of the common use of 'self-deception'. The situations in which we ascribe self-deception involve an expectation of knowledge, as we have argued earlier, on pages 5 and 6, but not knowledge itself.

The third type of error theory, and the final sort of philosophical account to be considered in this chapter, are what I shall call "Ought-to-know" Error Theories of self-deception. I shall be very brief, because this type of approach will be thoroughly discussed in the last chapter. There is only one account of this type in the literature, Siegler's,²⁰ and it is brief: I shall, however, be advocating a model of this type myself.

"Ought-to-know" error theories analyse self-deception as an unwarranted belief or cluster of beliefs about which the victim ought to know better. The key feature of such an account is the characterisation of the counter-belief as being within the self-deceiver's power, and to be expected given the circumstances. Neither motive nor intention is invoked to explain our intuition that self-deceivers are responsible for their ignorance, which differentiates this view from motivated error and intentional deceiver theories. Siegler treats 'self-deception' as a euphemism, but in my account the concepts of passive dissociation and mediate responsibility will be introduced to account for both our intuition of the self-deceiver's guilt, and the propriety of that intuition.

The next two chapters will examine the two most prevalent types of philosophical account in the literature or self-deception: motivated error theories, and intentional deceiver theories, respectively. An elaborated ought-to-know error theory will be developed in chapter five.

Notes to Chapter One

1. This is discussed later in this chapter and in chapter three.
2. An account of these states is given on p. 7.
3. See Hilary Putman, "The Meaning of 'Meaning'", in Mind, Language, and Reality, for an elaborate argument in a similar key.
4. F.A. Siegler, "Demos on Lying to Oneself", Journal of Philosophy 59, p. 473.
5. Siegler, p. 473.
6. Siegler, p. 474.
7. T. Penelhum, "Pleasure and Falsity", American Philosophical Quarterly vol. 1, no. 2, p. 259-60.
8. This could be fulfilled by outlining the bounds of the responsibility, or by demonstrating that there clearly is none, or that ambiguity is unavoidable.
9. Amelie O. Rorty, "Belief and Self-Deception", Inquiry 15, pp. 387-410.
10. Sartre argues that the paradox reappears at the level of the censor (Being and Nothingness, pp. 93-94). I am unsure, however, of the force of this argument.
11. D. Poff-Cliffton, Self-Deception, 1979.
12. Penelhum, p. 258.
13. Penelhum, p. 259.
14. Penelhum, p. 259. He says, however, that we may ascribe this knowledge on indirect grounds (p. 260).
15. Penelhum, p. 258.
16. Penelhum, p. 259.
17. Penelhum, p. 258.
18. Penelhum, p. 258.
19. Patrick Gardiner, "Error, Faith, and Self-Deception", Proceedings of the Aristotelian Society, 70, pp. 235-6.

CHAPTER II

MOTIVATED-ERROR THEORIES OF SELF-DECEPTION

Motivated-error theories are a diverse group: I would include Canfield and Gustafson,¹ Sartre,² and Poff-Clifton.

The first of these is primitive compared to the others, and has been more than adequately criticised. Canfield and Gustafson propose to avoid the paradox of self-deception by formulating a model analogous to self-command rather than deception of others. Self-command is represented as 'making oneself do something', which is alleged to amount to doing something in adverse circumstances. Self-deception becomes believing something in adverse circumstances, the adversity in this case being the conflict with the evidence.

Canfield and Gustafson

Canfield and Gustafson's reformulation of self-deception as a form of self-command is both inappropriate as an analogy and insufficient as an account. As Patrick Gardiner³ points out, the analogy to doing something in adverse circumstances, when transposed into the analysis of belief, seems rather more appropriate as a description of the holding of justified though uncomfortable beliefs than it is of self-deception. Furthermore, the Canfield-Gustafson model does not address the problem of just how one is to command oneself to believe something contrary to the evidence of which one is aware without conscious awareness of the fact that this is

exactly what one is doing, thereby "destroying the lie from behind" (Sartre's phrase).

The other two examples of this group to be considered here, the accounts of Sartre and Poff-Clifton, stress the role of interpretation in forming beliefs on the basis of raw evidence: there is an important connection here with the epistemology of the underdetermination of scientific theories.⁴ The idea here is that the justifiability of ignoring counter-evidence in some cases is in self-deception improperly invoked to serve some particular interest in believing a certain way.

I shall discuss each of the accounts mentioned at moderate length, and attempt to show that "motive", if construed as involving conscious, purposeful deliberation, reduces this sort of account to a kind of intentional-deceiver model, and that if it is not so construed, is objectionable for the same reasons as are the unconscious process theories. These theorists are characteristically vague on the question of how much deliberacy is involved in the operative "motives".

Sartre

Sartre's classic account of self-deception, which is set out in the chapter of Being and Nothingness entitled "Bad Faith" and in several of his plays, is an analysis in two parts, which appears to have been undertaken as much in the interest of providing support for his views on the nature of consciousness⁵ as it is in the pursuit of a coherent understanding of the phenomenon discussed. The account is

given to some extent within the context of his theory of consciousness, and a full treatment of Sartre's model would require a discussion of these views. I shall not attempt to do this here, however, partially because of the scope of what it would entail, and partially because the relevance of the theory of consciousness is primarily to the first part of Sartre's analysis, and I believe that the second part contains the essential features of the model as an account of self-deception.⁶ I shall, therefore, presume an understanding and an acceptance for the purposes of discussion of such Sartrean terminology as 'transcendence', 'facticity', and 'being for others' in my brief treatment of the first stage.

According to Sartre, the paradox of the apparent "lie to the self" requires resolution, when considered as it is in what he calls "pre-reflective consciousness", in terms of the dual nature of being - the transcendence and facticity of consciousness. Being can be correctly understood either from the aspect of what a person is -- his being-in-itself or facticity, or in terms of his freedom to become something else -- his being-for-itself, or transcendence.⁷ What is true about oneself from one of these aspects will almost always if not always be false from the other. Sartre claims that what may happen in self-deception is that a claim which is true in one sense will be advanced as true in the other context, and so in this sense the self-deceiver does

not make totally unjustified claims, although the strategy of making them in an inappropriate context remains unjustified. Sartre identifies the other dualities which supplement the basic account: that of being-for-oneself vs. being-for-others, that of being-in-the-world vs. being-in-the midst-of-the-world, and that of being-as-I-was vs. not-being-as-I-was.⁸ The first of these is fairly self-evident in its distinction. The second refers to the difference between taking oneself as an object and taking oneself as a person. The last refers to the senses in which one both is and is not the same person as one once was, after having transcended that past condition. In each case, a claim about one's nature can be made which is true in one mode of the duality and false in the other. In self-deception, claims true in one sense are advanced inappropriately in the other.

At this point in Sartre's account, we have been shown how a false justification might be produced for a desirable belief about one's nature, but we have yet to be shown how contrary evidence falls away, nor have we been given an indication as to how this strategy of false justification could escape the attention of a consciousness which is totally translucent⁹ to itself, nor why this might be undertaken. The second stage of Sartre's analysis attempts to provide answers to these questions.

The second stage of Sartre's model, operative in what he calls "reflective" consciousness, emphasizes the

similarity between self-deception and faith: hence 'bad faith', Sartre's term for self-deception. For Sartre, all knowledge involves faith: counter-evidence always exists, but beliefs adopted in good faith adhere to "the norms and criteria of truth as they are accepted by critical thought".¹⁰ In bad faith, however, there is an abandonment of these. Criteria for belief are lowered. There is a tendency "not to demand too much",¹¹ and a subtle shift from believing despite some counter-evidence to believing despite the bulk of the evidence being contrary.

The shift into bad faith is compared to falling asleep:¹² it is an attitude of flight caused by fear¹³ of being-in-itself. If we simplify this to "flight from reality", where the reality is the reality of one's nature, there is only a slight distortion of Sartre's view. Self-deception, for Sartre, is a perversion of normal belief management policies,¹⁴ induced by a fear of the reality of one's nature, such that a convenient belief about it may be held in an inappropriate context despite the awareness of counter-evidence.

Sartre's model can be criticised on at least two grounds.

First, it does not adequately cover the field of self-deceptive behaviour. Sartre specifies an attitude of flight due to fear as the only motive and sense of identity as the only object of self-deception. No justification is given, other than a selection of examples which have been

carefully designed to fit this mold. It seems to me that fear is not the only emotion associated with self-deceptive behaviour. One might, for example, deceive oneself out of pride, or desire for love. One of these would appear to be the most plausible sort of motive to associate with Siegler's Brown, cited on p. 8. A strategy of reducing these to fear seems possible here, but such an accommodation is not attempted by Sartre, and would be of dubious plausibility in any case.

Similarly, it also seems as if self-deception often has a feature of the world as its object, rather than always a feature of one's sense of identity. It would seem possible, for example, for someone to deceive himself about the nature of the institution in which he works. An employee of the post office might believe that the organization of which he is a part is committed to delivering the mail as quickly and as cost-effectively as possible, though it might seem inconceivable that someone in his position could be unaware of the fact that the majority of his co-workers were working at half-speed during regular hours in order to ensure that they would be paid double to clear up the backlog in over-time. In addition, it would seem that the object of Leonard's self-deception (page 1) is the psychological make up of Ronald Reagan, and that the object of Brown's self-deception (page 8) is his wife's behaviour. Once again, strategies of accommodation of these sorts of examples to Sartre's account might be

possible, but would not seem likely to be plausible and certainly not the most natural interpretation, and are not in any case attempted by Sartre.

The second, and principal objection to this account, concerns the mysteriously fear-inspired lowering of criteria for belief, which is the crucial feature of the second stage of Sartre's analysis. There are two, somewhat inter-related problems with such a formulation: questions of how we are to imagine that fear produces this result, and whether this production involves conscious deliberation.

Nothing more than the location of the process in conscious deliberation is required to make the action of a fear on processes of belief formation imaginable. But this is a measure which Sartre is unwilling to adopt,¹⁵ and with good reason: if the lowering of criteria is a conscious response to fear, the account is transposed to an intentional deceiver model, with all of the attendant difficulties that Sartre believes to be fatal to that model.¹⁶ Instead, the slip into self-deception is compared to falling asleep.¹⁷

But we cannot allow ourselves to be satisfied with an answer of this type, which postulates unconscious and indeterminate processes through which such thoroughly conscious phenomena as fear and belief formation are alleged to be causally connected, with no justification other than that the intentional deceiver model doesn't work. It seems fair to say that in spite of his displeasure with the psychoanalyst's "explanation by magic",¹⁸ he is himself

offering us an account that is equally magical, and as tempting a target for Occam's Razor. It would seem that the same objections apply to this interpretation of Sartre's fear-induced lowering of belief-formation criteria as to the unconscious process theories previously discussed: models which require unconscious processes to explain behaviour, in the absence of any independent evidence for their existence, are inferior in principle to models which do not.

We may say then that if the Sartre model is intended to involve a conscious fear-inspired lowering of criteria for belief, which seems rather unlikely, it will collapse into a type of intentional deceiver model. If on the other hand the model is taken to rest on unconscious processes, as seems far more plausible it will be subject to the objections of these.

Poff-Clifton

The second of the motivated-error theories that I shall consider is that of Deborah Poff-Clifton, in her Self-Deception, an MA thesis produced at Carleton University. Poff-Clifton's model draws on the notion of circumstantial evidence:¹⁹ she makes an explicit connection²⁰ between the view that scientific theories are underdetermined by empirical data and her own claim that beliefs may be underdetermined by the evidence relevant to their adoption, and still be rationally held.²¹ The difference between normal and self-deceptive policies of belief-management appears to amount to how much counter-evidence is admitted,²² though this is not entirely clear. The feature of motivation in her model may

be intended to aid in specifying this distinction as well as be an essential part of her explanation of self-deception.

Three strategies of self-deceptive belief-management are listed, along with their non-self-deceptive correlates. These are:

- (i) pseudo faith vs. faith
- (ii) pseudo-implications-blindness vs. implication-blindness
- (iii) pseudo-rational accommodation of contrary evidence vs. rational accommodation of contrary evidence.

I shall briefly outline what appears to be intended in each of these distinctions:

- (i) Pseudo-faith is alleged to differ from true faith in its operative motive. Both appear to be forms of motivated distortion of belief-management, however. Pseudo-faith is an acceptance of a belief unjustified by the evidence due to self-interested motives, whereas faith is an acceptance of unjustified belief due to altruistic motives.²³
- (ii) Implication blindness is a condition which is to occur when the drawing of a certain implication would threaten foundational beliefs. Pseudo-implication-blindness is a variant of this referring to cases in which the beliefs threatened are not truly foundational, but merely cherished.²⁴

(iii) Rational accommodation of contrary evidence is the acceptance or rationalization of contrary evidence in cases where there are good reasons to believe that the initial belief is true. Pseudo-rational accommodation is the same procedure in cases where there are strong, self-interested motives for believing the initial belief true.²⁵ This distinction is reminiscent of Sartre's good and bad faith, at the level of reflective policy, the differences being Sartre's specifications of fear as the motive and the self as the object. Poff-Clifton's generality seems more plausible to me in this instance.

There are a number of problems with this account. The first two are specifically Poff-Clifton's, and the last is the general objection to theories of this type.

First, it would appear that Poff-Clifton considers faith, in her sense, to be a normal and acceptable kind of belief-management strategy. But I fail to see the relevant difference between holding an unjustified belief altruistically and holding it self-interestedly. Both seem equally self-deceptive. This is not to say that her distinction between faith and pseudo-faith is entirely misguided: her account of 'faith' may be closer to what is ordinarily meant by the term than is Sartre's "good faith". But it is mistaken to attempt to distinguish faith from self-deception in this way. If self-deception is to be characterized as a motivated

departure from sound epistemic policies, then it would appear that believing something evidentially unsubstantiated for altruistic reasons would also be self-deceptive. The distinction between faith and pseudo-faith is thus of questionable relevance to this, or any other model of self-deception.

Incidentally, I should say that I have a preference for Sartre's account of faith. We will recall that for Sartre the difference between good faith and bad is determined by good faith's adherence to sound epistemic policies. But bad faith is also a kind of faith, and an appropriate way to distinguish religious faith from the good faith of a Galileo. (the application is mine, not Sartre's).

It is even more difficult to see the purpose in the distinction of implication-blindness and pseudo-implication-blindness, once again because it seems that neither is a normal and acceptable belief management strategy. There is, in addition an oddness to considering them strategies at all.

If the "blindness" described is a strategy, albeit an ill-named one, then it must amount to a deliberate policy of refusal to draw intolerable or unwelcome conclusion. But however, sound this sort of thing may be in a practical sense, it will surely always be unsound as an epistemic policy, regardless of how damaging the avoided conclusion might have been if drawn. Self-deception need not always be an impractical strategy: the legitimacy, not the prudence of belief is what is at issue. Consequently, the distinction of pseudo and true

implication blindness is irrelevant if these are construed as strategies of belief management.

An additional problem: if implication blindness is a strategy, does this not reduce the account to one of the intentional deceiver type?

It would seem more natural,²⁶ however, to treat the sort of phenomenon that Poff-Clifton calls 'implication-blindness' as blindness: a mere inability to draw certain conclusions. Once again, however, it proves difficult to make a distinction between true and pseudo-implication-blindness that is relevant to self-deception. If we are prepared to allow a characterisation of self-deception as an involuntary blindness, an approach to which I have no immediate objection in principle,²⁷ we are in doing so adopting a model of the self-deceiver as victim rather than perpetrator. The distinction between pseudo and true implications blindness, if a dysfunction rather than a strategy, would have to be made in terms of the strength of the threat, objectively determined, to the emotional well-being of the victim. But what force can there be in such a distinction where there is no question of avoiding the error, as the blindness is involuntary? To say that one sort is just fine and the other self-deceptive would be arbitrary indeed, as the victim is as innocent and pure as the driven snow in either case.

Regardless of how we take Poff-Clifton's concept of implication blindness, the difference between the true and pseudo forms of it appears irrelevant. Furthermore, the inclusion of even an undifferentiated concept of implication blindness would seem hazardous for an error theorist, as it introduces a danger of reduction to either intentional or non-intentional deceiver theory.

The last of Poff-Clifton's strategies of self-deception, "pseudo-rational accommodation of contrary evidence" -- which I shall call PRACE for short -- appears to differ from its rational correlate (RACE) in the substitution of a self-interested ²⁸ motivation for believing in the truth of the initial belief in the place of good reasons to believe in the truth of it; and, as previously noted, it resembles the Sartrean distinction of good and bad faith at the reflective level.

This argument is paradigmatic of motivated error theories: self-deception appears as a dysfunction of normal epistemic policies of hostile-evidence management caused by the action of motives. The question that we must ask, and which PRACE and other motivated error theories characteristically gloss over, is this: how are we to construe the causal effect of the motive? The word 'motive' would suggest that it acts as a desired objective in conscious deliberation, but if so, the motivated error theory will collapse into the intentional-deceiver model that error theorists allege to be hopelessly paradoxical. If, on the other hand, 'motive' is to be

understood in a special non-deliberative sense, then its action becomes mysterious and requires explanation. An account of the nature of this action is not to be found in Poff-Clifton's text, but it would seem that postulation of some sort of unconscious processes would be required, and if this were to be the case, the same charge of vulnerability to Occam's Razor as was made against the other theories involving unconscious processes would be appropriate.

Our survey of motivated error theories has shown that these models are radically defective. In all three of the theories analysed, the attempt to avoid the allegedly paradoxical intentional deceiver model by substitution of a concept of motive-distorted belief-formation strategies reduces either to the intentional deceiver model, or to an unconscious process model, when 'motive' is fleshed out.

Notes to Chapter Two

1. Canfield and Gustafson, "Self-Deception", Analysis 23.
2. Jean-Paul Sartre, Being and Nothingness, p. 86-116.
3. Gardiner, p. 228.
4. This connection is explicitly noted by Poff-Clifton.
5. See Sartre's remarks on p. 116, Being and Nothingness.
6. Sartre's own assessment, see p. 112, Being and Nothingness.
7. Sartre, pp. 98-99.
8. Sartre, p. 100 ff. These are introduced on p. 100 and discussed thereafter.
9. Sartre, p. 89.
10. Sartre, p. 113.
11. Sartre, p. 113.
12. Sartre, p. 115.
13. Sartre, p. 116.
14. I have adapted this from Poff-Clifton's term, 'Belief Management Strategy'.
15. Sartre, p. 113. The sense in which a Sartrean self-deceiver is responsible for his state does not involve intention, but rather neglected self-improvement. Fingarette contends that Sartre strongly condemns the self-deceiver as a deliberate scoundrel (Herbert Fingarette, Self-Deception, London: 1969, p. 137 ff), but he appears unable to give references, and I cannot find any either.
16. Sartre, p. 87-90.
17. Sartre, p. 113.
18. Sartre, p. 95.
19. Poff-Clifton, p. 137 ff.
20. Poff-Clifton, p. 153.
21. Poff-Clifton, p. 154.

22. Poff-Clifton, pp. 156-157.
23. Poff-Clifton, pp. 161-163.
24. Poff-Clifton, pp. 164-167.
25. Poff-Clifton, pp. 167-174.
26. Poff-Clifton, however, appears committed to the first interpretation, though I cannot be sure. Her text contains conflicting statements, (i.e. p. 137 paragraphs one and three).
27. Though I would have such an objection if the blindness turned out to be dependent on unsubstantiated unconscious processes, which seems likely.
28. I shall not belabour the irrelevancy of this qualification.

CHAPTER III

INTENTIONAL DECEIVER THORIES AND THE CONCEPT OF ATTENTIONAL FOCUS

The best known and most influential type of philosophical theory of self-deception is that which I have called the intentional deceiver type. Philosophers to be counted in this group include Demos,¹ whose article "Lying to Oneself" marked the first interest in self-deception in recent analytical philosophy, and Fingarette,² whose book Self-Deception is still the most sustained treatment of the issue to date. As stated in the first chapter, intentional deceiver theories hold to a basic claim that self-deceivers deliberately bring it about that they are deceived. The first problem with which such a theory must come to terms is what has been called the paradox of self-deception: the question of how one and the same person can be both the perpetrator and the victim of deception. It would seem that if one is to deceive someone about something, one must be aware of the truth in order to be able to lead that someone into error, and it would also seem self-evident that if one is to be said to be deceived about something, it cannot be the case that one is aware of the truth of the matter. At least, this is certainly the way it is in the deception of others.

Intentional deceiver theories attempt to avoid this "paradox" -- it is not really paradoxical, but rather wildly

implausible per se as an account of human cognition -- by means of the concept of attentional focus. They explain the problematical "knowing ignorance" with reference to the scope of a person's immediate attention as distinguished from the wider field of available knowledge. The picture is familiar: concentration on a foreground item without regard to other consciously available information, as when one is able to concentrate exclusively on a game of chess in a room polluted with party talk. The attempt to assimilate the paradox through this concept is usually as follows: the conflicting beliefs, or the beliefs which conflict with the available evidence, can be rationally and simultaneously held so long as no more than one is attended to at a time.

The basic model is usually elaborated to include the claim in self-deception (as opposed, presumably, to delusion), immediate attention is deliberately focused away from belief or evidence which conflicts with a particular cherished belief. The insistence on intentional attentional focus is prompted by concern for conceptual criteria with which to distinguish self-deception from concepts such as stupidity and obstinacy and wishful thinking, and the legitimacy of this concern we have questioned in the first two chapters. But the principal awkwardness into which these sorts of theories fall is distinct from this, and concerns the fine structure of the supposed selection process. It would seem that this selection process, if it is to be "intentional", must itself take place

in immediate awareness. But if the selection is always to be "in focus", some question must arise as to how one can be said to be unaware in the end of evidence which was so recently weighed, found undesirable, and rejected. What would seem to be required, if this were to be possible, would be an implausible **degree** of opacity of consciousness to itself, such that what one was thinking about just a moment ago might now be inaccessible.

Resolution of this problem has been approached via the notion of levels of description.³ It has been argued that if contrary evidence were rejected under a general description, intentionality might be preserved without necessitating a postulation of opacity of consciousness to itself in order to avoid reinstatement of the paradox. On this view, it would be possible to reject a general description of the contrary evidence consciously, without the necessity of appraising it in detail. We shall examine Fingarette's detailed version of this approach in detail shortly. But in general we may say, as David Kipp notes, that there is a "blurring generality of the theory's own focus"⁴ on a number of crucial questions, notably that of deliberacy. Rejection on the basis of a general description is supposed to allow the retention of intentionality in the process of selection: "but, if he has taken no account at all of the relevant specific aspects ... it is hard to see how he can be charged with anything other than ignorance or oversight regarding those aspects".⁵

FINGARETTE ON THE MECHANISM OF SELF-DECEPTION

The most elaborate and sustained intention-deceiver analysis of the mechanism of self-deception is to be found in Herbert Fingarette's book, Self-Deception (1969). Fingarette's treatment is an attentional focus theory which relies heavily on an active model of explicit consciousness as the exercise of a skill. Much is made of a new "active" terminology, but if retranslated back into the more usual and "passive" terms of 'belief' and 'awareness', Fingarette's account assumes a distinct similarity to that of David Pugmire. Whether or not something essential would be lost in such a retranslation -- this matter will be addressed further on in this section -- Fingarette is undoubtedly more ambitious than Pugmire (whose paper "'Strong' Self-Deception" also appeared in 1969). Pugmire's formulation of the "generalized description" version of IAF theory was intended to account for a "strong" form of self-deception that arises "only after the weight of the evidence for the unwanted conclusion has finally collapsed the stubborn defenses against it" (Pugmire, p. 345), and which is "the last step short of derangement" (p. 346). Mr. Fingarette, on the other hand, intends to develop an account of self-deception generally which will "unravel the paradoxes", clarify the language associated with self-deception, and also "bend familiar usage" in order to "isolate, identify, and discuss certain features of the self-deception situation with which there is no everyday language associated" (all from Fingarette, p. 5).

"Were a portrait of man to be drawn, one in which there would be highlighted whatever is most human, be it noble or ignoble, we should surely place well in the foreground man's enormous capacity for self-deception." (Fingarette, p. 1)

What Fingarette does is to analyse self-deception as a result of the purposeful misuse of an attentional skill, such that certain evidence is not made explicit when it would be appropriate for this to be done. A fairly elaborate explanation is given as to how this might be possible, and it is with this section of his account (section III of his book) that I shall be dealing, almost exclusively. I shall not be concerned here with Fingarette's investigations into why self-deception occurs and is so prevalent, though I suspect that his claim that all cases of it are caused by a person's attempts to preserve a particular sense of identity at all costs is mistaken. This is not to say that the issues of how and why are entirely separate; it is rather that I think that the flaws in Fingarette's account of how self-deception is possible can be quite adequately dealt with in isolation, and that once this has been done there is no need to follow the account any further.

I shall first outline Fingarette's proposed explanation of the mechanism of self-deception, then pass on to consider the accuracy and significance of his new "active" terminology, the peculiar "tacit" sense of 'awareness' that he employs, and his claim that self-deception is a "self-covering policy". I shall make references throughout to the

criticisms advanced by Bela Szabados,⁶ Ronald de Sousa,⁷ and Alan Drengson.⁸ In the end, I hope to show that Fingarette's major difficulties stem from the familiar and quixotic requirement of IAF theory, that the deception be both deliberate and unconscious.

Fingarette proposed an "active model" of consciousness, to be counterposed against what he styles as the received, "passive" model of consciousness, in which explicit awareness is taken as given, and anything less than that is a material subtraction in need of causal explanation. He notes an analogy in this model which extends the imagery of visual perception to the conceptual framework of belief and awareness in general.⁹ Information pours in as from the visual field, unless obstructed by some "blind spot" (whether or not this is an adequate characterisation of perception is of course open to debate). Fingarette suggests that this "visual" model is inappropriate for conceptual analysis of consciousness, and although perhaps not entirely dispensible in the end, the passive "cognition-perception"¹⁰ terminology ought to make way for a new analysis in "volition-action"¹¹ terms. What this amounts to for Fingarette is a reversal of polarity in descriptions of consciousness. Explicit consciousness is no longer to be taken for granted; rather, "explicitation" will be an act, or a series of them -- Fingarette calls this "spelling-out".¹² This is to be understood as "an activity which has a close relation and analogy to linguistic activity" (Fingarette, p. 39), although he cannot specify

the exact relationship (p. 40). No matter, for the important thing is the new model, the analogy. "Spelling-out" is also to be understood as a skill, the proficiency in the exercise of which varies widely amongst individuals (p. 39).

An important aspect of the skill of "spelling-out" is that although the exercise of it is ubiquitous and not something that one pays attention to -- one does not ordinarily "spell-out" spelling-out -- it is nonetheless a rationally determined activity. Explicit ~~consciousness~~ is the exercise of a particular skill for a specific purpose; one must decide whether or not to spell-out by weighing the reasons for and against doing so (p. 42). For the most part, these assessments are "tacit", but they may also be spelled-out in many cases -- though not that of self-deception.

"Engagement" is Fingarette's term for the object of an instance of spelling-out. He says that this is not a precise technical term, but rather a convenience used to refer to "what someone does or what he undergoes as a human subject" (p. 40):

"It is a matter of the activities he engages in, the projects he takes on, the way the world presents itself to him to be seen, heard, felt, enjoyed, feared, or otherwise 'experienced' by him ... What a person does not somehow take account of is not part of his engagement in the world." (p. 40-41)

I suspect that what Fingarette means by 'engagement' -- though as Szabados has noted¹³ he leaves the term crucially vague -- is a way of being situated in the world rather than an item of one's phenomenal experience, and also that it is

expressive primarily of particular situations in his use of it, rather than of situatedness as a whole (as his introduction of the concept quoted above would seem to indicate). Fingarette does equate 'his engagement in the world' with 'his world', and throughout his discussion of it the term is used only in the singular. Nonetheless, I believe that this is merely unfortunate and misleading phrasing, since in every subsequent use of the term, an "engagement" is one of many. I am further persuaded that nothing grandiose is intended in it by Fingarette's nonchalance in adopting it. The first point -- that an "engagement" is a kind of situation rather than an item of phenomenal experience -- is more difficult to demonstrate without an extended discussion that I do not want to get into here, since the issue is not particularly relevant to the kinds of criticisms that I want to make of Fingarette's account. But briefly, it seems to me that an "engagement" is presented as a kind of transaction that a human subject has with the external world, of which he need not be conscious explicitly,¹⁴ but with which he is involved in some active way. Significantly, self-deception may often involve erroneous spelling-out of an engagement, so on Fingarette's account an engagement must be something that one can be mistaken about. It has an objective core.¹⁵ This is, of course, hardly a "knock-down" argument, and so the backing for my view will have to be its own plausibility. But I shall suggest that

if 'engagement' is taken as a kind of weasel-word for 'belief', or a generic term of which 'belief' would be an appropriate instance, then Fingarette will have been seriously misrepresented.

Another aspect of Fingarette's account which requires some discussion is the peculiarly restricted sort of intentionality¹⁶ that he considers to be attributable to an individual's acts of self-deception. In the later sections of his book it becomes evident that Fingarette does not believe it appropriate to hold someone morally responsible for his self-deceptions, or for his subsequent handling of the engagements about which he has deceived himself (p. 140-41).

"There is thus in self-deception a genuine subversion of personal agency and, for this reason in turn, a subversion of moral capacity ... There emerges the 'medical' view of self-deception as helplessness due to 'mental pathology'." (p. 141)

"If we turn .. to a consideration of the role of the moralist, ideologue, or minister, we can see now that insofar as they presume the self-deceiver to be fully a personal agent in the matter at hand, they preach, teach, and argue in vain. The futility of preaching to the neurotic has been remarked by the psychiatrically oriented (p. 143)

It is nonetheless abundantly clear in the early chapters that Fingarette considers self-deception to be a complex volitional act, involving reasons, judgements, and policy (e.g. p. 42). It is further to be noted that Fingarette makes it his special task to develop a non-paradoxical way of talking about the essential (for him) oddity of self-

deception, the aspect of "knowing, intentional ignorance" (p. 29), and that his primary criticism of the previous literature was of the failure to take this into account.

What emerges from all of this is a view of self-deception as an intentional act which once executed subverts the normal relationship between agency and responsibility.

Whether or not an engagement is spelled-out is, as previously mentioned, dependent on whether or not there is a reason for one to do so. There may also be reasons not to spell-out an engagement. And in cases of particularly noxious engagements, says Fingarette, a policy of not spelling them out may develop,¹⁷ or be adopted¹⁸ (I shall allow the distinction between these two to remain unexamined for now). Now, as I have also discussed previously, spelling-out is a form of engagement,¹⁹ and as such is susceptible of being spelled-out. This is not, however, ordinarily done.²⁰ In the cases in which a policy of not spelling-out has developed, moreover, the preliminary and tacit assessments leading to the decision not to spell-out, or to adhere to the policy, are never themselves spelled out, for to do so would be to destroy the effectiveness of the policy. A policy of not spelling-out some particularly noxious engagement is hence a self-covering policy²¹ -- its very nature rules out the spelling of it out. This is the basic account of self-deception,²² in terms of how it is possible, and it is this account of the mechanism of deception which contributed to the development of IAF *

* - Intentional attentional focus

theory. A self-covering policy of avoidance of spelling an engagement out even when appropriate is self-deception.

I shall review and consolidate a little before proceeding with analysis and criticism of Fingarette's account. The "focusing" imagery acquires a slightly different sense with Fingarette -- it is no longer a restrictive focus ("focusing away from"), but rather a resolving focus²³ (in the same way that a photographer "focuses in on" his subject), a procedure coextensive with what is more commonly thought of as explicit awareness, and which in the case of self-deception is not undertaken even though apparently appropriate.²⁴ If we spite Fingarette and reformulate his view in "cognition-perception" terms, it would seem that we might say that the details of the noxious and odious engagement about which one is deceiving oneself are allowed to remain in a general -- or better, a "diffuse" state of awareness, in a situation in which it would be appropriate to make them explicit. That is, in a situation in which the formation or correction of beliefs about what is the case in the world is or ought to be involved.²⁵

That an engagement is left inexplicit is deliberate, and in this respect Fingarette sounds like Demos and Pugmire. It is true that one "decides" to allow the engagement to so remain; nonetheless, there is a distinction between this account and that of Pugmire beyond that of scope, insofar as Pugmire's self-deceiver is supposed to be trying to accommodate his conflicting beliefs by "formulating" the

malodorous one in general euphemistic language. Fingarette does consider similar cases, and it is fairly clear, I think, that he would consider such "forumulating" as a form of spelling-out.²⁶ The "decision" that Fingarette's self-deceiver makes regarding the details of an execrescent engagement is made tacitly, which is to say, without spelling-out. However, Fingarette's account does share with that of Pugmire the explanatory feature of the insulation of the self-deceiver's explicit awareness from unwelcome evidence by means of generality of description. It is just that the -- polarity? -- of the explanation has been reversed, and there is no longer any need to do anything to a belief (or awareness of an engagement) in order to make it general or diffuse -- I simply omit making it explicit.

I now propose to examine two critical areas of Fingarette's analysis which I feel are seriously flawed. Fingarette wants to introduce a new model of consciousness, and new terminology to describe it, which will analyse consciousness in terms of action and volition rather than in terms of cognition and perception. He wants to make talk of "knowledge" and "belief" irrelevant to the analysis,²⁷ and in my view that is not possible, as will be argued shortly. Self-deception is essentially a cognitive phenomenon. But even if we can overcome this difficulty, perhaps by taking Fingarette's action terminology and model as supplementary explanatory devices useful in dealing with the cognitive

(which is plausible), there are irremediable difficulties with the account of self-deception that Fingarette gives within his framework. The model of explicit consciousness as a skill which is only executed when a tacit assessment (or judgement) of the reasons for and against doing so has come up heads, and the characterization of self-deception as a tacit policy of tacitly assessing not to make certain engagements explicit, are both given inadequate justification based on inappropriate analogies. The model of self-deception in particular seems a strict impossibility. I shall proceed to expand on these criticisms in order of introduction.

According to Fingarette's opening pages of the chapter dealing with the mechanisms of self-deception, previous analyses of self-deception have dealt with the problem "primarily in terms of belief and knowledge, or in terms of 'perception' language such as 'appear' and 'see'" (p. 34). Fingarette proposed that this "cognition-perception family" of terms is inappropriate to describe consciousness (p. 35), and that another "family", the "volition-action family", can do better. The term 'conscious' and its variants are to switch allegiance, and become "volition-action" terms, and ultimately there is to be a "volition-action" account of self-deception. (p. 35) The chief difference is to be the elimination of "passive, visual imagery" in cognition-perception terminology and the replacement of it with "a different but reasonably coherent set of imagery, metaphor, and idiom" (p. 38) in which explicit consciousness will be an activity or skill.

A number of justifications are offered for this change of terminology and model of consciousness. Fingarette appears to believe that the passive visual metaphors of the cognitive-perceptive terminology are responsible for the paradoxicality of self-deception (p. 34), or that the only way of describing self-deception in that terminology involves hinting by means of paradox (p. 64, p. 5). As Szabados has noted (Szabados, p. 22), oblique reference is made to an ability of the new model to explain self-deceptive attitudes, wishes, and fears,²⁸ which the old model allegedly could not account for. But Fingarette's primary argument is that the new model is more successful, and the new imagery more fertile (p. 37); and to that one can only say that we shall have to wait and see.

As de Sousa has noted (de Sousa, p. 313-14), the exact extent of Fingarette's proposal is left unclear -- Fingarette does not entirely abstain from the use of the old model,²⁹ and in any case states quite plainly on p. 35 that he does not intend to "eliminate the 'cognition-perception' family of terms from the analysis of self-deception". But then on p. 64 he says:

"Therefore, if we wish to put matters directly and non-paradoxically, instead of indirectly and paradoxically, we must turn to this largely unexplored way of characterising consciousness instead of to the familiar knowledge-belief-perception approach".

This sounds a good deal less conciliatory. Since one would naturally want to avoid paradox if possible, just what role is there for 'knowledge', 'belief', and 'perception' in this

new account of self-deception? None, really. I suspect that on p. 35 Fingarette is merely trying to soften the blow. On p. 63, and in discussing sincerity, he is determined enough to try to eliminate the terminology of the cognition-perception model from the formulation of the problem of self-deception. Instead of having someone who intentionally misleads himself, we now have "purposeful discrepancy between the way the individual really is engaged in the world and the story he tells himself" (p. 63). Everyday usages of 'know' and 'believe' regarding such behaviour are analysed as crude references to the engagement and the story, respectively (p. 63-4). The new model of consciousness seems intended to completely supplant the old, in accounts of self-deception at least.

Can the cognitive problem be so easily explained away? One may doubt, with Drengson (p. 479), whether a convincing analysis of consciousness can be given which does not involve belief, but it is certainly much more questionable whether such an analysis of consciousness could generate an account of self-deception. The central feature of the philosophical problem is, as Fingarette himself notes, the element of the appearance of "knowing, intentional ignorance" (p. 29); it will at least be necessary to relate the new model back to the cognitive framework of what is to be explained. Fingarette is conscious of this, and tries to develop points of contact between his own account and common usage,³⁰ but the difficulties are not merely linguistic.

Fingarette proposes to replace the self-induced false belief of the original formulation with "telling oneself a false story" (p. 63). The deceived state is one in which a person does not spell out the actual engagement, but rather an amalgam of the non-threatening aspects of the actual engagement and a "cover story"³¹ which fills in the gaps. The difficulty with this reformulation is that one can imagine³² cases in which a person might tell himself such a story without believing it true, and also cases in which a person "parades"³³ an accurate account of an engagement before himself and others without actually accepting it as true, without believing it. It would seem more natural to consider the latter of these to be a case of self-deception, but Fingarette seems committed to saying that only the first is, even though the person is not "deceived"! This divergence is rather more than is acceptable in a new model, I fear. It would seem that it is not enough, then, to tell oneself a false story -- if one is to be said to be deceived one must also believe it. Similarly, that one spells out an engagement just as it is, is no guarantee that the true story is accepted as true, and so may well be a form of self-deceptive diversion. Fingarette is not entirely unaware of this problem -- his extended discussion (p. 50-4) of 'sincerity' may be taken as an attempt to address this, amongst other issues. This, however, serves best to demonstrate that the problem with Fingarette's reformulation is embedded in his conception of sincerity. I quote his final position:

"We do have a typical use of 'sincere' in which the ascription of sincerity is not intended to imply that what a person tells us about his engagement conforms to the actual nature of his engagement.

What we do always mean to imply when we ascribe sincerity is a conformity between what the individual tells us and what he tells himself." (p. 52)

Criteria of ascription are not a full account of usage. No doubt we do ascribe sincerity on the basis of the sort of conformity that Fingarette mentions. but what we mean when we say that a fellow is sincere in what he says is that he believes what he says. Normally, of course, what a man tells himself and what he believes to be true coincide; but it is important to remember when they part company, sincerity rides with belief.

Fingarette's attempted reformulation of the account of the self-deceiver's deceivedness in non-cognitive terms is thus unsuccessful. When reference to belief is abjured, it is no longer possible to effectively differentiate the deceived state from the fully apprised -- and more pointedly, it becomes difficult to assign a precise meaning to 'deceived'. It would therefore seem that cognitive terms cannot be dispensed with. In fact, I shall now try to show that in his reformulation of the sense in which the self-deceiver "knows what he is doing", Fingarette cannot help surreptitiously reintroducing belief.

I have discussed earlier in this section (and I shall later belabour) the fact that for Fingarette the intentionality of self-deception inheres in the tacit

assessments of whether or not to spell-out an engagement, which in the case of self-deception is always (or almost always -- in deference to some of the foregoing discussion) a decision not to spell-out, in adherence to a tacit policy of not spelling that engagement out. It is only reasonable to suppose that these tacit assessments also correspond to the sense in which a self-deceiver was said to know of his self-deception in the old way of talking. And insofar as we suppose that the self-deceptive policy would be relatively consistent, the only sense in which a self-deceiver would know of the details of the engagement about which he is deceiving himself would be the extent to which they entered into the reasonings of his tacit assessments of whether or not to spell them out. So it would seem that for Fingarette, one's "knowledge" of the policy of self-deception, and of the details of the engagements about which one self-deceives, is confined to tacit assessment.

The deciding of whether or not to spell-out involves consideration of reasons for and against doing so, and a weighing of these where there is conflict. It would be difficult to imagine how this could be done, tacitly or otherwise, if the process did not involve factual beliefs about the nature of the engagement, the implications of the awareness of the details of that engagement, and the relative values of these implications in terms of one's own best interest -- or in terms of whatever value is to be the scale. What are we to make, then, of Fingarette's "non-cognitive" analysis, which

must have recourse to belief at the crucial point in the application of his key concept (spelling-out) to the basic problem of intentionality? That this recourse is alleged to be -- mysteriously! -- "tacit", hardly helps.

In summary of this minor section, it seems that Fingarette's "action-volition" model of consciousness and of self-deception is intended to supplant the use of the older "cognition-perception" model, and that new terminology is meant to replace cognitive terminology such as 'knowledge', 'belief', and 'perception'. But the reformulations have not proved adequate, and Fingarette's own account seems to involve surreptitious use of cognitive concepts. Self-deception would seem to be an irreducibly cognitive problem. It remains quite possible, however, that Fingarette's new model may serve a useful supplementary role in providing a solution to the cognitive problem. It is to that question that I now turn.

I discussed earlier in this section (p. 6) Fingarette's view of self-deception as a rational and intentional act of a sort for which one is not morally responsible. One puts oneself into self-deception, but one cannot avoid it on the occasion,³⁴ and so one is also its victim. Fingarette accomplishes this by characterising the decision that his self-deceiver makes as to whether or not to spell-out as a "tacit" one, which is a part of a "tacit self-covering policy". My analysis will show that the analogies that Fingarette uses to justify his dependence on these concepts are inappropriate, and that the concepts themselves -- particularly the "tacit self-covering policy" -- are incoherent.

Essential to the model of explicit consciousness as a skill, and perhaps the primary reason for Fingarette's conceptualization of spelling-out as a skill rather than just as an activity, is the assessment stage preliminary to the actual execution of the skill.

"Exercise of the skill requires sizing up the situation in order to assess whether there is adequate reason for spelling-out the engagement ... Where it happens that there are both reasons for spelling-out and reasons for not spelling-out the very same thing, these are presumably weighed against one another by the individual. Whether or not he spells-out the matter in question, or exactly how he does so if he does spell it out, will depend not only on his assessment but also upon his ingenuity in adapting to the conflicting considerations. All this is no more than we allow to be the case with complex skills generally". (p.42-3)

Because it is also true that in most cases, and certainly where self-deception is concerned, such assessments are not themselves spelled-out (p. 43), the intentionality of decisions not to spell-out is limited to this tacit level of assessment. It is therefore Fingarette's position that a fairly complex and intentional reasoning can take place with no more than tacit awareness of the process. He supports this, and attempts to explicitate the sense of 'awareness', through a number of examples. The tacit assessment alleged to precede spelling-out is compared to the assessments involved in deciding when to execute the various procedures associated with the skill of driving a car, and with assessments of when to speak, what to say, etc., associated with the skill of using language:

"Skill in driving a car is not merely the ability to perform certain movements; it is also the ability to assess the possibly conflicting considerations in the situation in order to settle which movements to perform, when to do so, and precisely how. Skill in speech calls for assessing just when to speak, how to speak, what to say. Skill in spelling-out requires analogous assessments". (p. 43)

One cannot help feeling a bit surprised upon finding that this is the only justification that Fingarette sees fit to provide. One is inclined, at first, to be nasty about it, for Fingarette's examples are suspiciously inappropriate, particularly the driving skill example. It is, after all, fairly obvious that the execution of a driving skill, or even a linguistic skill, is different in kind from the relevant act of assessment in a way that the execution of spelling-out is not. Stepping on a brake and telling someone where to get off are both motor sequences, when considered apart from the assessments of when, whether, and how to execute them; the assessments, on the other hand, are thoughts. Now it is evident that one may think about stepping on the brake without actually doing so, but is it quite as evident that one may think about thinking about a personal failure without thinking about it in the process? On the face of it, Fingarette's clarification by analogy would seem to fail rather dismally.

But perhaps what Fingarette is trying to do here is something a little less implausible than the bare invocation of an analogy. The assessments that we make in driving, and in speaking, are typically very quick and sure, in marked contrast to the sort of laborious concentration involved in,

say, adding up the monthly bills, or in trying to analyse why one can't skate. These skill assessments are for the most part almost automatic, although in situations in which considerations conflict, or something unexpected happens, one usually pays more attention to them.³⁵ Usually, though, one is not particularly aware of these assessments, and Fingarette can be profitably taken to be arguing that decisions about whether or not to spell-out an engagement are normally of this type. "Weak" forms of awareness are discussed by Fingarette (p. 45-6), but he does not -- and again one is suspicious of this omission -- attempt to specify the sort of "weak" awareness applicable to tacit assessment. But one of the senses that he discusses seems close:

"Also contrasting with spelling-out is another 'weaker' sense of 'conscious' in: 'Are you conscious that you are shuffling the cards?' 'Yes, of course, I'm perfectly conscious that I'm doing it, but that hasn't distracted me from what you are saying; I'm paying attention only to what you say and nothing else enters my mind.'" (p. 45)

The shuffling of cards is presumably a skill of the relevant sort, which is to say that it involves decision as well as motor sequences, and so I suspect that this is the sort of awareness that Fingarette imagines we have of our tacit assessments of whether or not to spell-out. That this sort of weak awareness can normally be spelled-out on demand is no obstacle to the identification of it with tacit assessment, for Fingarette holds this to be true of

tacit assessments as well (p. 43 -- because it is a component of the skill of spelling-out, which can itself be spelled-out). If this is Fingarette's intention, then we have located a sense in which actions of which a person is only vaguely conscious can be executed. The trouble is, though, that we do not yet have a model in which rational decisions are arrived at tacitly; the shuffling of cards in this automatic way is the execution of a decision to shuffle that was probably quite explicitly conscious. Similarly in the case of driving skill, there are decisions -- such as whether to turn at the next corner -- but these are conscious. The assessment of how far to turn the wheel in order to make that corner would not be conscious -- it would, if you like, be tacit. But it is a long hop from this, which might as easily be called a conditioned response as an assessment, to the kind of decision procedure described on Fingarette's p. 42 and quoted on my p. 16.

In short, though it may be possible to tease out some idea as to what sort of awareness Fingarette thinks we have of our decisions regarding spelling-out, the concept of a "tacit decision" doesn't make sense intuitively, and Fingarette fails to make sense out of it. The analogies he offers are not appropriate, and no other justification is offered. However, since I suspect that some sense might be made of this if Fingarette's over-rationalization of his concept were attenuated to some degree, I shall proceed to

examine the second major concept, the "tacit self-covering policy", even though this depends entirely on the notion of tacit assessment.

For Fingarette, self-deception is not characterised by isolated refusals to spell-out an engagement, but by a persistent tendency to do so (p. 47), which according to him is to be explained as the result of a deliberately adopted policy, which is also tacit (p. 48-9). Drengson (p. 479) asks why Fingarette should ever have bothered with such a concept, as it seems to him to be unnecessary, since no engagement that there was a reason to avoid spelling-out would ever survive tacit assessment. Nor, I shall add in support of Drengson, is there any reason to fear that the tacit assessment not to spell-out might itself be spelled-out for some reason or by some chance, since the same reasons not to spell-out the original engagement would be reasons not to spell out the tacit assessment. So it is a puzzle why Fingarette should choose to incorporate this concept into his account. Though I cannot support my suspicion, I think it likely that Fingarette does so at least in part so as to reinforce the intentionality of self-deception in his account.³⁶

In any case, the policy which Fingarette's self-deceiver adopts is necessarily a "self-covering" one. This means that the policy of not spelling-out a particular engagement also involves not spelling out the policy, since to do so would involve spelling-out the engagement, which would frustrate the purpose of the policy (p. 49).

There are two points that I wish to make about this, and I shall begin directly with the first. One would think that a policy, and especially adherence to a policy, would be an inherently explicitly conscious thing.³⁷ Fingarette characterises the inability of the self-deceiver to spell-out his engagement as analogous to the case of someone constrained by a promise:

"This inability to spell-out is not a lack of skill or strength; it is the adherence to a policy (tacitly) adopted. 'He cannot admit it, cannot let himself become conscious of it', here means 'He will not'; but the 'will not' refers to a general policy commitment and not an ad hoc decision not to spell it out. There is a parallel to: 'It's not that I won't come with you, but that I can't because I promised Smith I'd go with him'. The can't (sic) of self-deception is markedly different in one respect from the 'can't' of the man who promised. The promiser can't say yes but he can say that he is committed not to say yes. The self-deceiver, however, is committed to say nothing on either score". (p. 47-8)

Adhering to a promise, even for Fingarette, it would appear, is an explicitly conscious policy, the very vigilance of which ensures success. But does this enlighten us as to how one is to adhere to a policy, i.e. the self-covering policy of not spelling-out, that one is not to be explicitly conscious of? It is another case of misleading analogy, and in this case Fingarette is further misled (or further misleads) by the constant use of speech terminology to make points about a phenomenon of consciousness, trading on the ambiguity of 'spelling-out'. For example, he says of the self-deceiver that he "is committed to say nothing on either score" (p. 48), meaning presumably that he is committed to think nothing

explicitly on either score. But the analogy does not go through. A self-covering policy of never saying 'cow' (for example) would be possible: this would mean only that one consciously adopt a policy of not saying 'cow', and that one would take care not to mention the policy specifically. One would not say, "I've adopted a policy of not saying 'cow'." This would be rather odd, but it could be done. In the self-deception version, however, what would be required in a self-covering policy would be that I not explicitly think about whatever it was that I wished to deceive myself about, but also that I never explicitly think about my policy of not explicitly thinking about it either. If all of this is to be required, then one may well ask what it could mean to say that the self-deceiver is committed to such a policy: it cannot mean what is meant when we say that a promise-maker commits himself to keep his promise, or what it means when we say that someone has committed himself never to say 'cow', for in both of these cases what is meant is irreducibly deliberative.

Briefly, I have shown in my discussion of Fingarette that the "action-volition" model of consciousness that he proposed is not an adequate replacement for the traditional "cognition-perception" model, especially with regard to self-deception. Fingarette's own account involves surreptitious use of the concepts and terminology it is intended to supplant.

In the specific account of self-deception, Fingarette's treatment of the key concepts of "tacit assessment" and of the "self-covering policy" of self-deception are flawed by his reliance on inappropriate analogies. Fingarette fails to support his concept of "tacit assessment" with any examples of decision-making which is not "explicitly conscious" (in his sense of this term), nor has he given any clear sense to the initially implausible idea of a self-covering policy of not spelling-out. Insofar as the device of "tacit assessment" was the means by which intentionality was to be included as a feature of the account without disruption of the self-deceiver's state of ignorance, Fingarette has failed to show that self-deception is in any sense deliberate, and as such his account misses its own objective, which was to explain the "knowing, intentional ignorance" which Fingarette claims in his first chapter as the essential feature to be explained.

The difficulty in accommodating both intentionality and genuine ignorance as features of self-deception are not, as we have seen, merely Fingarette's: they are endemic to the whole project of intentional-deceiver theory. Having outlined the basic form of intentional-deceiver theories and of their use of attentional focus as an explanatory device, and having demonstrated how the most sophisticated exponent of this type of account has failed to avoid its characteristic pitfall, I shall now consider a different sort of attentional focus theory, which attempts to avoid the problem by dropping the intentionality feature altogether.

RORTY ON THE MECHANISM OF SELF-DECEPTION

The last of the accounts to be considered in this section, which is expounded by Amelie O. Rorty in her article entitled "Belief and Self-Deception", spans two of the categories of our classification. It shares with the motivated error theories discussed in chapter two, the explanation of self-deception in terms of a distortion due to the influence of motives of normal belief - management strategies, but her account of the mechanism responsible for this distortion invokes a modified form of the concept of attentional focus which is characteristic of intentional-deceiver theories. Since these categories were not intended to be rigid or exclusive, as I noted in the first chapter, I decided that in view of its structural similarities to intentional deceiver theories, Rorty's account would be most perspicuously dealt with here. Rorty makes an attempt to deal with the most trenchant difficulty afflicting explanations of self-deception in terms of attentional focus, and so it would seem reasonable to treat her theory as a successor to those of Fingarette, Pugmire and Demos.

Rorty's basic strategy can be seen in the two following quotations:

It is because assertions of beliefs are used to perform a wide range of speech acts besides that of speaking the truth, and because there is a large area of indeterminacy in such assertions, that self-deception is possible.³⁸

It is because indeterminate truths and declarations can be spoken with one's eye resolutely on their general appropriateness, keeping the details necessary for their specification or enactment out of focus, that we can believe and not believe, mean and not fully mean what we say, even in circumstances when it might be important to have the details in mind.³⁹ (Rorty's emphasis)

To paraphrase crudely, Rorty would appear to be arguing essentially this: given that beliefs do not (or at least, that assertions of beliefs do not)⁴⁰ always express what one considers factual, and given that beliefs (or assertions of beliefs) are often indeterminate, it is possible to entertain beliefs or make assertions of belief on the grounds of a general appropriateness, ignoring the details of the context in which they are appropriate, and that this is the explanation of self-deception. There would appear to be three points in this which are particularly in need of explication: what is meant by 'indeterminacy', the relevance of the "wide range of speech acts", and the nature of the mechanism of focusing employed. We shall examine each of these in turn, as they arise in our exposition of Rorty's account.

According to Rorty, there are two senses in which assertions of beliefs may be indeterminate with regard to relevant evidence: first, beliefs are not always truth-positing, and second, truth-positing beliefs are often held in spite of not being fully justified by the relevant evidence. Two types of truth-positing belief which may be rationally held without adequate justification are identified:⁴¹ pre-critical

beliefs, which are those acquired largely by accretion (acculturation, conditioning) without ever having been subjected to rational analysis, and those which, though evidently lacking adequate justification, are held in the hope that such evidence will be forthcoming, or because they are useful and not entirely implausible, or because there is no alternative which is substantially more plausible (i.e. scientific contexts). In addition, she identified a kind of non-truth-positing belief, pragmatic belief, which comprises those given tacit assent out of convenience in the absence of definitive evidence as to their truth or falsity.⁴²

The significance of all this is that Rorty wishes to argue that what happens in self-deception is that the strategies we use accommodate important beliefs with relatively insignificant counter-evidence, and normal non-truth-positing uses of beliefs (or assertions thereof), go awry: beliefs admissible as hypotheses are accorded factual status,⁴³ beliefs appropriate only pragmatically are treated as if rationally justified.⁴⁴

What this would appear to amount to is that Rorty is claiming that a person may commit epistemic errors of the following types:

- (i) holding a belief that is unreasonable on the grounds that it would be reasonable to hold it without complete justification,

- (ii) treating a justified inference as unjustified because it is not absolutely certain, even though far more likely than any alternative account of the evidence,
- (iii) holding a belief in a truth-positing context which is appropriate only in a non-positing context.

The first two of these would appear to be similar to the sorts of epistemic errors identified by Poff-Clifton and Sartre. The last is specific to Rorty. What is now required is an account of how a person could come to commit such epistemic errors, and so it is to Rorty's account of the mechanism of self-deception⁴⁵ that we now turn.

Self-deception often involves unexpected and problematic focusing: a person seems to be ignoring what is most salient in a situation ... Declarative sentences that normally convey beliefs ... are used to bring about some end other than 'telling it like it is'. It is essential for the successful performance of the non-veridical speech act, that the declarative sentence is also used as a bearer of truth-value, though it is not used to bear its proper truth-value.⁴⁶

In these first moves of Rorty's account of the mechanism of self-deception two important aspects of it can be discerned: the application of speech-act analysis of other-deceptive assertions to self-deception, and the introduction of the concept of attentional focus. This last is as we have seen common currency amongst intentional deceiver theories. We shall consider the application of this speech act analysis to self-deception immediately below, and

- (ii) treating a justified inference as unjustified because it is not absolutely certain, even though far more likely than any alternative account of the evidence,
- (iii) holding a belief in a truth-positing context which is appropriate only in a non-positing context.

The first two of these would appear to be similar to the sorts of epistemic errors identified by Poff-Clifton and Sartre. The last is specific to Rorty. What is now required is an account of how a person could come to commit such epistemic errors, and so it is to Rorty's account of the mechanism of self-deception⁴⁵ that we now turn.

Self-deception often involves unexpected and problematic focusing: a person seems to be ignoring what is most salient in a situation ... Declarative sentences that normally convey beliefs ... are used to bring about some end other than 'telling it like it is'. It is essential for the successful performance of the non-veridical speech act, that the declarative sentence is also used as a bearer of truth-value, though it is not used to bear its proper truth-value.⁴⁶

In these first moves of Rorty's account of the mechanism of self-deception two important aspects of it can be discerned: the application of speech-act analysis of other-deceptive assertions to self-deception, and the introduction of the concept of attentional focus. This last is as we have seen common currently amongst intentional deceiver theories. We shall consider the application of this speech act analysis to self-deception immediately below, and

attempt to determine the legitimacy of Rorty's analogizing between the cases of other-deceptive assertions and self-deception, proceeding thereafter to attempt to determine whether her account is any more plausible if treated as a contribution to conventional attentional focus theory.

An other-deceptive speech act, or more generally a non-veridical speech act, is said to be dependent -- if it is to be successful -- on the normal use of assertions to express beliefs. Non-veridical speech acts, and especially other-deceptive ones, usually consist of assertions which do not express what is believed to be the case, but are used as if they did. The deception of others would, as Rorty appears to be arguing, be difficult if not impossible were there not a preponderance of cases in which assertions are used to express beliefs about what is believed to be the case; and it is certainly the case that it becomes rather more difficult to make effective non-veridical use of one's assertions if one comes to be suspected of doing so regularly. It is indeed, as Rorty claims, specifically because the sincerity of assertions is not ordinarily challenged unless the other is given good reason to do so that the deception of others can be effective.

Her argument depends on the claim that there is an important and revealing analogy between the functionings of other-deceptive assertions and self-deceptive beliefs. Underlying this claim and similar arguments is a fairly plausible doctrine which treats the concepts and grammar of thought as

dependent upon, generated by (and perhaps none other than) the concepts and grammar of language. Ms. Rorty's attempt to analyse the essentially cognitive process of self-deception isomorphically with a form of analysis of other-deception is a natural extension of that doctrine for which the strategy of understanding beliefs in terms of assertions is probably a necessary requirement. The doctrine of logical isomorphism would not seem to imply of itself, however, that thought and language be isomorphic in practice as well, and particularly not in a special case such as that of self-deception. The attempt to establish the existence of such a practical isomorphism is nonetheless legitimate and not unmanageably implausible, and this is what I take the relevance of the speech-act analysis to be: an explanation by analogy.

It would seem that what Rorty must be arguing is that the analysis of the mental act of self-deception, by analogy to the speech-act of other-deception, ought to proceed under the assumption that in the majority of cases -- and certainly in the paradigm case -- beliefs which are entertained in thought are sincerely believed to be true; that is, they may or may not be justified beliefs, but they are not entertained despite a conscious, and sincere conviction that they are at variance with the truth. In the majority of cases, a belief entertained in thought can be considered to be entertained factually.⁴⁷ In a large number of cases the justification of a belief need not be co-present

in consciousness (p. 389-90), and these beliefs are, as it were, given the "benefit of the doubt" and treated as if true. Indeterminate beliefs are most often susceptible of justification if challenged, or are at least almost always sincerely held. It is because of this, Ms. Rorty would appear to suggest that self-deception is possible. In self-deception, a proposition is entertained as if it were a factual belief although it is not in fact this sort of belief. An example of an entertainment which would not be a factual belief would be something one wanted to believe, another would be hypothesis, yet another something entertained "for the sake of argument". (p. 389-91).

So it goes with the analogy to deception of others. The difficulty of course is that it is not all that easy to imagine how one might mistake a wishful thought or hypothesis for the real truth-positing thing, in the deliberate and systematic way characteristic of self-deception. The analogy with other-deception breaks down here, because my ability to deceive another is dependent upon his having no direct access to the mode of my assertion (or to be more exact, my 'declarative'). If it is to be claimed that I deceive myself in a similar manner, what will be required is that I can have no direct access to the mode of entertainment of my own belief. This is not immediately plausible: are we really to think, for example, that when I entertain a wishful thought I may remain ignorant that it is entertained wishfully, such that I might mistake it for a truth-positing belief? It would be a serious drawback to Rorty's

account if it were required simply to postulate that consciousness is opaque to itself with regard to the mode of entertainment without any explanation as to why this might be. An interpretation of this opacity which immediately suggests itself is as a sort of intentional "turning away" from the mode of entertainment, which would place Rorty clearly as an intentional deceiver theorist. I shall argue, however, that although she does adopt the concept of attentional focus in her account of this opacity, her theory is not really of the intentional deceiver type.

The second aspect of Rorty's account of the mechanism of self-deception, the use of the concept of attentional focus, is best illustrated in the second of the quotations with which we opened our discussion. For the sake of easy reference, I shall cite it again:

It is because indeterminate truths and declarations can be spoken with one's eye resolutely on their general appropriateness, keeping the details necessary for their specification on enactment out of focus, that we can believe and not believe, mean and not fully mean what we say, even in circumstances when it might be important to have the details in mind.³⁹ (Rorty's emphasis)

There is a marked similarity between this account and those of intentional deceiver theorists such as Pugmire and Fingarette, who attempt to accommodate the concept of an intentional "focusing attention away from" evidence contrary to a cherished belief by the condition that the self-deceiver remain unaware of the evidence to be avoided, through the invocation of "levels of description". This approach, discussed earlier, depends on the plausibility of claiming

that if contrary evidence were deliberately rejected under a general description, intentionality might be preserved without necessitating either the postulation of an improbable opacity of consciousness to itself or the reinstatement of the paradox. On this view it would be argued that it is plausible to construe the rejection of hostile data under a general description as deliberate, even though no detailed appraisal of the hostile nature of that data would have been involved. The flaws inherent in this type of account have been discussed at length earlier in this chapter: it was demonstrated that the claim of intentionality could not be maintained if no account at all is taken of the specific aspects of the data hostile to the self-deceiver's belief, and that without the concept of intentional avoidance no account at all is being given of the holding of the oddly unjustifiable beliefs characteristic of self-deception. Rorty substitutes general appropriateness for generality of description as the level at which evidence can be unconsciously and yet in a sense deliberately selected.

The account, incorporated as an attentional focus theory, would appear to run like this: there are a number of modes in which a statement or entertainment can be appropriate, and it is possible and usual in casual thought for an individual to choose (in what must be a very restricted non-cognitive sense of "choosing" -- Rorty doesn't elaborate) to reject or to accept the statement or entertainment as

appropriate or inappropriate without specifying details. If accepted, it may function in ensuing deliberations without further attention to the context of its appropriateness. Deliberate selection, on the basis of appropriateness is thereby alleged to be possible without requiring awareness of the details of the potential entertainment, and without specification of the mode to which it is appropriate.

We can identify a modified version of the analogy drawn earlier by Rorty between the speech act analysis of other-deceptive assertions and self-deceptive entertainments: an entertainment appropriate in some non-veridical context can be accepted as generally appropriate and used as if it were appropriate as a factual belief. In this modified version, however, that possibility is dependent on the indeterminacy of belief, with regard to justification, discussed earlier. The fact that beliefs are often indeterminate with regard to the sense within which they are justified, together with Rorty's claim that in a majority of cases no error is made in accepting such beliefs without such specification,⁴⁸ would appear to mitigate the implausibility of the simpler version to some extent, insofar as the modified account does not appear to require an unexplained and implausible degree of opacity of consciousness to itself. Unfortunately, these two aspects of Rorty's account do not quite match up: she has shown that beliefs are characteristically indeterminate with regard to their justification, but what the modified account requires is indeterminacy regarding

the context of appropriateness of a belief, and this has not been demonstrated.

The difference between indeterminacy regarding justification and indeterminacy regarding context of appropriateness would appear to me to be this: the former refers to the holding of beliefs which are not rationally justified by the evidence available to the believer, whereas Rorty's sense of the latter would refer, if it referred at all, to the entertaining of beliefs as generally appropriate for entertainment without specification of the mode of entertainment (i.e. wishing, hypothesizing, truth-positing, etc.). The former is justifiable in terms of Rorty's appeal to the indeterminacy of belief with regard to justification, but the latter appears neither to be plausible nor to have been argued for. What would be required would be either an explanation as to how an individual might fail to have access to the modes of thought in which he is concurrently engaged or evidence that such a lack of access exists in fact, but to this end Rorty offers nothing beyond the analogy to speech act analyses of other-deceptive assertions found to be inadequate in our consideration of the simple version of her account. As noted before, the idea that I might entertain a wishful thought and mistake it for a truth-positing belief is not particularly plausible.

Rorty's attempt to produce a non-intentional attentional focus account of the mechanism of self-deception fails by relying on an initially doubtful and ultimately unsubstantiated analogy

between assertion and thought, much as Fingarette's failed by relying on analogies which were actually misleading. We shall conclude, then, that the concept of attentional focus is unsuited to the development of an account of the mechanism of self-deception. Such an account is required, and in the fourth chapter I will approach the problem of giving such an account in a new way.

Notes to Chapter Three

1. Raphael Demos, "Lying to Oneself", Journal of Philosophy 57, pp. 588-595.
2. Herbert Fingarette, Self-Deception, London 1969.
3. For example, Pugmire, "'Strong' Self-Deception", Inquiry 12, pp. 399-346. Fingarette's account is also of this type.
4. David Kipp, "On Self-Deception", Philosophical Quarterly 30, p. 310.
5. Kipp, p. 311.
6. Béla Szabados, "Fingarette on Self-Deception", Philosophical Papers 6.
7. Ronald de Sousa, "Review of Self-Deception", Inquiry 13.
8. Alan Drengson, "Critical Notice of H. Fingarette, Self-Deception", Canadian Journal of Philosophy vol. 3, no. 3.
9. Fingarette, p. 35.
10. Ibid.
11. Ibid, p. 36.
12. Ibid, pp. 34-38.
13. Szabados, p. 24. Szabados's reconstruction, however, is a misconstrual, and his criticism on this point misconceived. The mother's expressed belief that her son is a good boy is part of a separate engagement. There is no paradox, as Szabados suggests there is. One engagement is partially spelled out, another is not.
14. Whether any sort of consciousness is necessary is just not clear.
15. It may be that an oversight of this point was responsible for Szabados's confusion, note 13.
16. Drengson (p. 480) takes Fingarette to be claiming that self-deception is not intentional but merely purposeful, and is understandably perplexed about the distinction. But this is not Fingarette's position, for he was 'intentional' and 'purposeful' interchangeably throughout, as in "We are mainly concerned with the more complex kind of self-deception which is intentional, purposeful". (Fingarette, p. 52).

17. Fingarette, p. 47.
18. Ibid, p. 49.
19. Ibid, p. 43.
20. Ibid, p. 41.
21. Ibid, p. 49.
22. Ibid, p. 62.
23. Though Fingarette does not like visual metaphors, this one seems appropriate.
24. Fingarette, p. 47.
25. Drengson presses the question of what 'appropriate' means here (Drengson, p. 479), suggesting that self-deception may often be most "appropriate". He is too unsympathetic, as Fingarette is clearly referring to cases in which truth is sought and the material relevant.
26. Fingarette, p. 41 (the sailboat example). Some critics have failed to note the difference, e.g. David Kipp, p. 310-11.
27. Actually, there is some doubt what he is trying to do, supplant or supplement. This is discussed in the next few pages.
28. But the old model can easily deal with these, as Szabados shows (Szabados, p. 22).
29. For example, p. 47, where Fingarette is trying to distinguish authentic spelling-out from the behaviour of an actor "playing" spelling-out. Also, see Fingarette p. 63 (top), and my discussion on pp. 79-80 of this chapter.
30. For example, Fingarette p. 5 and p. 64.
31. Fingarette, p. 50.
32. One need not imagine, as they are among the most common strategies. Szabados (p. 24), de Sousa (p. 315), and Drengson (p. 478) each bring this point up in their own ways. Szabados feels that this possibility erodes the distinction between self and other-deception in Fingarette's account. But Fingarette is careful to specify that the false story be the only story that the subject tells himself (e.g. p. 52), which undermines Szabados. Drengson suggests that such a story need also be avowed, and that this casts doubt on Fingarette's view that avowal must precede spelling-out: Drengson claims that it is possible to spell-out now, and disavow later. But as the sort of story under discussion in Fingarette is an account of an engagement as mine, I do not think that avowal can really be at issue here.

33. de Sousa, p. 314. de Sousa is concerned with the criteria of authentic spelling-out.
34. Fingarette (p. 47-8) says of this, that it is not a lack of skill or strength, but adherence to policy. But if so, he must still account for the adoption of the policy, a decision also reached tacitly (p. 48), and of the sense of 'adherence'.
35. But not in the sense of 'spelling-out', which is necessarily quasi-linguistic. For Fingarette, all senses of 'conscious' other than spelling-out are "weak" (p. 45-6). It is interesting to note that amongst the myriad weaker senses of 'conscious' is the sense in which I would be conscious of a red-hot coal dropped unexpectedly into my bare hand.
36. Fingarette seems sensitive about the possibility that the self-deceiver might not be capable of spelling-out his engagement, e.g. p. 47.
37. Drengson notes this, but does not elaborate. (Drengson, p. 479).
38. Rorty, p. 387.
39. Rorty, p. 392.
40. Rorty believes talk of assertions to be interchangeable with talk of beliefs, p. 388.
41. Rorty, p. 389-390.
42. Rorty, p. 389-390.
43. Rorty, p. 392.
44. Rorty, p. 402.
45. This is what she calls it, p. 388.
46. Rorty, p. 388.
47. I am using the word 'entertain' to refer to the appearance in consciousness of a belief. This term like 'declarative', is undetermined regarding mode.
48. Rorty, p. 389-90.

CHAPTER IV

MOOD-DISSOCIATION AND SELF-DECEPTION

In this section I will examine some effects that emotions have on episodic memory. Self-deception almost in an account of the causal structure of self-deception (if not always) involves strong emotions. Most philosophical analyses of self-deception have given these emotions the role of motivational factors, which, as it were, "bribe" one into abandoning the structures of intellectual conscience. Such a motivational effect would appear to be unavoidably conscious, and one to which one's attention would be drawn; this is typical of goal-oriented behaviour in general, and it is this model that the motivationalist is drawing on. The problem, of course, is that the particulars of self-deception are such as to make awareness of the motivating factors impossible without defeating the execution of the project. Motivationalists -- and I include virtually every philosopher who has ever considered the problem of self-deception -- have tried out various strategies of accommodation, but these have tended to slip into paradoxical and implausible accounts of consciousness. The choice of the philosopher on this last point has usually seemed to have been determined by his views or the morality of self-deception, rather than on any evidence as to whether or not deliberate agency is actually a feature of self-deception, as if the former could be settled

"by intuition" without reference to the latter, being prior to it.

In the account that develops I shall consider an emotion to be a complex psychological state (or activity) with both cognitive and physiological aspects. There are certainly characteristic judgements of fact and value, but there is also an underivable and powerful physiological determination of behaviour associated with emotion. I shall not attempt to arbitrate here as to whether or not one or other of these is the right method of approaching the analysis of human emotions. What I propose to do is to take a look at some recent research in psychology which may be taken to suggest that changes in mood state have significant effects on cognition, not by means of the influence of some variant of the libido concept, or other "demon of desire", but rather through determination of the informational field in which cognitive processes operate. The concentration, then, will be on the effects on cognitive activity of what is presumably a physiological state, though predominantly behaviorally defined in the experiments cited.¹ I shall not attempt to define 'mood', 'state', or 'mood state' very exactly, since on my view the philosophically proper terms would have to be physiochemical, and at the present neither I nor anyone else is equipped to do this.

I hope to be able to show that strong emotions have significant extra-motivational effects on cognitive activity. I will attempt to demonstrate that these effects are sufficient

to account for the phenomena of self-deception.

Mood State-Dependency Effects

In a paper² given at a recent meeting of the American Psychological Association as a Distinguished Scientific Contributions Award address, Gordon Bower presented a growing body of evidence of significant "mood congruity" in human cognition. In a series of experiments involving the manipulation of mood through hypnotic suggestion, Bower and his associates found that material learned in one mood was best recalled in the same mood, that mood had a definite effect on free association, that ambiguous or neutral information tended to be interpreted and elaborated in a manner consistent with the mood state of the subject, and that non-mood-ambiguous information tended to be better remembered if consequent with the mood state of the subject at the time of presentation.

No doubt each of these effects might seem of interest in a discussion of self-deception in terms of the effect of emotion on cognition, but for several reasons I am going to concentrate on mood state-dependent retention in episodic memory.⁴ First of these is that this effect is comparatively well documented; mood-state dependency of episodic memory has been reported by a number of research groups using various experimental methodologies.⁵ There is a

further connection to a much longer literature about drug-state-dependent episodic memory.⁶ A second reason is that it seems possible that the other "mood congruity" effects might ultimately be explainable in terms of state-dependency of retention. If it could be shown that the field of memory information to which one has access at a given time was determined to some extent by one's emotional state at that time, it would not seem surprising that ambiguous or neutral stimuli might be slantedly interpreted, given that the field of experiences that might be accessed for purposes of comparison must have substantial influence on the outcome of interpretive processes. Lastly, it would seem to me that there is in the concept of mood-state-dependency of episodic memory a potential to increase the content of our explanations of self-deceptive behaviour, and a particular relevance to the question of agency. The claim that interpretations and judgments tend to be polarized by emotional states is both plausible and familiar, but this sort of data is too undeterminate with regard to structure, almost all theories would have to take this sort of "mood congruity" into account to be even intuitively plausible. But the incorporation of mood effects on memory into an account of self-deception allows an unpacking of vague talk about the influences of desire and emotion on cognitive judgement, through the identification of fairly specific effects of emotion on a specific constituent of cognitive judgement -- the field of accessible memory.

Furthermore, whereas mood-effects on interpretation and free associations are also indeterminate with regard to intentionality (a consequence of their indeterminacy with regard to structure), a mood state dependency of recall is already an effect at the point of consciousness of the successful recall or unsuccessful attempt.

Having given a brief outline of why I think this matter relevant to self-deception, I shall now proceed to expand on the concept of state-dependent-recall, and to review some of the experiments which have demonstrated it. I shall start with an account of what it means for recall to be state-dependent, using a drug experiment to demonstrate this. I shall then examine the extension of this paradigm to the effect of mood-states, and I shall review some of the experimental confirmations of this. I shall then explore the consequences that this would seem to have for psychological models of memory.

The term "state-dependent-learning" identifies an experimental methodology, a set of findings, and a conceptualization of these results as a general property of discrete states of consciousness,⁶ specifically mood-states or drug-states.⁷

As it does in animal studies, the SDL phenomenon refers to both a set of operations and a set of findings. Although these operations have become increasingly complex (Overton, 1974), the basic design unit of approach and the germinal findings remain as follows. A set of behaviors or events is learned or stored in one state, defined either in behavioral, pharmacological, neurochemical, or inferred terms, and is later tested or retrieved in either the same or in some disparate state.

A subject learns (or stores, encodes, or inputs) some information while in a particular drug-defined state and later attempts to utilize (output) the information either while in that same (congruent) state or in a different (disparate) state. In general, information utilization is impaired when the subject's state is changed between the input and output phases of the experiment relative to the situation in which state is held constant for input and output. This pattern of results -- or, more specifically, the interaction of drug state at input with drug state at output -- defines the phenomenon of SDL.⁸

The sort of experimental design that is typical of SDL experiments is the 2 X 2 factorial design, in which performance at a given task, previously taught is measured for two experimental groups under two different sets of conditions. An adapted representation of data obtained by Bustamente et al⁹ is given below to demonstrate the form of a 2 X 2

factorial experimental design showing state dependent retention. Some subjects were shown a slide of 8 geometrical shapes while under the influence of either amphetamine or amobarbital, and others viewed this slide after having ingested a placebo. The subjects were then subsequently tested two days later under the opposite condition -- those who had learned under the drugs were asked to recall after having ingested a placebo, and vice-versa. Then two days later, the subjects were tested again in the same state as that in which they had originally learned the shapes. The results graphed here are evidence of a significant dissociation of learning. The subjects were best able to recall the shapes in the congruent state, even though this was tested twice as long after the presentation of the shapes as was disparate state recall.

Analogous results have been obtained in studies involving a wide variety of drugs which act on the CNS, and notably with both marijuana¹⁰ and alcohol.¹¹ (Drugs which do not affect the CNS, do not seem to produce SDL, even though they may have discriminable peripheral sensory effects.¹²) The effects of most of these drugs are such that they do not inhibit learning or retention in themselves, but only when alternated with a placebo or disparate drug state in the encoding and recall phases. That is to say, that learning and retention are in most cases¹³ as good in a DRUG-DRUG encoding-recall situation as in a PLACEBO-PLACEBO encoding-recall situation. The SDL effect therefore appears to be one of dissociation; material learned in one state tends to be less accessible, or accessible not at all, in another.

These effects have been well replicated¹⁴ and the dissociation interpretation has withstood substantial methodological criticism, discussed above and in some of the footnotes.

Recently the state-dependent-learning paradigm has been extended to investigations into the effects of mood-states on learning and memory. The same sort of dissociation appears in experiments in which mood-state is the manipulated variable; and since it is likely that the alterations of awareness which constitute (or at least cause or correlate with) mood states are connected with (and perhaps even caused by) the release and restrictions of physio-chemical substances, the extension is both intuitively plausible and confirmed in practice by a number of investigations manipulating mood in different ways:¹⁵ I shall discuss three of these experiments.

The basic MSDR effect is well demonstrated by Bower's 1978 experiment, which is reported in the 1981 paper previously cited. Highly hypnotizable subjects were instructed under hypnosis to adopt the required moods.¹⁶ They were then taught a list of sixteen words, after which they were instructed to alter their mood, and were then taught a different sixteen word list. When the subjects were later asked to recall the lists either in a congruent or disparate mood state, the following results were obtained, clearly demonstrating an SDL-effect.

Figure 2. Percentage retention scores depending on the match between learning mood and recall mood. The sloping lines refer to subjects who learned the two lists under different moods.

132 • FEBRUARY 1981 • AMERICAN PSYCHOLOGIST

In experiments with bipolar manic-depressive patients, Weingartner et al (1977) also found a state dependency effect. This experiment is worthy of special attention for two reasons: firstly, because it shows a gradual effect dependent on the degree of mood-disparity, and secondly, because they found that the SDL effect was especially pronounced when the affected information was sequentially related.

In the graph below, a graduated effect can be seen holding between change in mood-state and accuracy of memory. Weingartner and his associates used a sophisticated system of mood assessment involving 19 different scales. The subjects were asked to free-associate a list of words while in one state, and even later re-tested on their memory of these in different states.

Figure 20.3. Absolute change in manic state-effect and its relationship to retrieval in self-generated learning. Nine subjects were studied twice in each treatment condition. The average subjective-state changes experienced just before learning information is plotted as a function of retrieval of stored information 6 hr later. Procedures used to study learning and measures of subjective state change are described in detail elsewhere (Henry *et al.*, 1973). Copyright 1973 by the American Psychological Association.

This finding is similar to some degree to the incomplete experiment reported by Bower (1981), the results of an attempted correlation of strength of dissociation with absolute difference between moods, using Plutchik's categorization of similar and dissimilar emotions. These results are shown below

		RECALL MOOD			
		FEAR	ANGER	JOY	SADNESS
LEARNING MOOD	FEAR	.88	.52	—	.68
	ANGER	.44	.88	.77	—
	JOY	.53	—	.78	.88
	SADNESS	—	.73	.88	.78

Figure 9. Matrix of percentage retention scores for a list learned in the mood shown in the row and tested in the mood shown in the column. The same-mood conditions on the main diagonal are shaded; the opposite-mood conditions are circled.

2 2
OF / DE

That a gradient of this type could be established would be significant in establishing some sort of direct, and likely causal relationship between change in mood and field of memory. Bower and Weingartner were attempting to discover qualitatively different continua; Weingartner varied the intensity of the absolute difference between two disparate moods with reasonably exact clinical definitions and criteria, whereas Bower attempted to plot a continuum¹⁷ of the familiar types of mood in terms of the relative strengths of the dissociations between them, according to Plutchik's theoretical model (below). These are not, of course, conflicting continua; if one were to accept Plutchik's model, one might interpret Weingartner's data as a plotting of the relationship between "joy" and "sadness" of differing intensities in bipolar patients. The second

Figure 8. The eight primary emotions (inner labels) and their adjacent blends (outer labels) obtained by factor analysis of similarity judgments of emotional terms. (From "A Language for the Emotions" by R. Plutchik, *Psychology Today*, February 1980, p. 75. Copyright 1980 by Ziff-Davis Publishing Company. Reprinted by permission.)

noteworthy feature of Weingartner's study is his discussion of evidence indicating that what is lost in the disparate retrieval state is the sequential organization and information that cements events together in episodic memory.¹⁸ A number of investigations have noted that tasks involving sequential recall are particularly sensitive to disparate state dissociations.¹⁹ Weingartner and

his associates in the 1977 study found that amongst the words successfully recalled by their subjects, congruent-state recall was far more likely to preserve the sequence of words in the encoding phase, in addition to being far more accurate.

A comparison of the results of experiments involving sequential recall vs. those which did not is presented in Weingartner's article in Drug Discrimination and State-Dependent Learning.²⁰ It was found that of experiments using sequential recall or some variant of it, 83% showed a disparate state dissociation effect. Of the experiments which, according to Weingartner, did not involve memory for temporal relations amongst the to-be-retrieved events, only 15% showed state-dependency. This would seem to be fairly compelling evidence of a specific effect on sequential memory, were it not that the experimental types which, for the purposes of this comparison, are posited as paradigmatic of tasks involving (respectively) sequential and non-sequential memory, do not seem to me to pick these features out successfully. Specifically, the "representative paradigms" used for non-sequential tasks are item recognition and paired associate learning. It is certainly true that studies of these designs, and particularly the item recognition design,²¹ have on the whole failed to show any state dependency effect of significance. The trouble is that one cannot distinguish whether this result is due to the asequentiality of these tasks, or to the fact that they are both also paradigms of cued recall -- another type of memory task which has frequently been shown to be

rather unsusceptible to dissociative effects.²² This might seem to be a surprising oversight, since Weingartner has reported the effect of cuing in earlier experiments, and also two pages further on in this very article, but I suspect that his view²³ is that self-generated cues are stored sequentially, and that the overcoming of dissociation in cued-recall experiments amounts to producing coding information which would be far more likely to be independently producible were it not for the impairment of sequential memory caused by the disparate state condition. This may well be the case, but what is important here is that the paradigms of comparison are confounded with another variable, namely, the effect of cuing. Consequently, pending an explicit demonstration that self-generation of cues involves sequential memory in a fundamental way, the suggestive results of this comparison must be put aside; though it is still an interesting speculation, to which we will briefly return.

The results of the earlier studies, mentioned above in the last paragraph but one, are not confounded in this way. In view of this type of result, it is fair to conclude that memory for sequence appears to be particularly disrupted by disparity of state, even though these results must as yet be regarded as preliminary.

The importance of this apparent effect of disruption of sequential memory to the current enterprise is that sequential searching would seem to be a principal means of episodic memory retrieval. Of course, speculation of the type found in the

preceding paragraph is not intended as any substitute for empirical confirmation of the hypothesis that a person's memory of episodes of his personal history is subject to disparate state dissociation. This sort of data is available, and is to be presented forthwith. The intention is, rather, simply to indicate the consonance of current speculating about the fine structure of the dissociation phenomena with the hypothesis that memory of personal episodes would show this effect.

Evidence that mood dissociation does affect recall of events of personal history, and other meaningful information, was obtained by Bower, and reported in Bower (1981). His subjects showed significant, though somewhat assymetrical,²⁴ mood state-dependent dissociation effects in each of a series of related experiments.

In the first experiment,²⁵ subjects were asked to keep a daily diary of emotional incidents for one week. After an interval of a further week, the subjects were hypnotized, and random halves were put into happy and unhappy moods, and asked to recall as much of the diary as possible. The percentages of recall are shown in the graph below. Note that in either recall mood, congruent state incidents were more likely to be recalled than opposite state incidents, although the disparity is greater in the "pleasant" recall mood, and small recall percentages are also lower in this state. This last aspect

Figure 3. Percentage recall of incidents from an emotional diary depending on whether the incident was pleasant or unpleasant and whether the subject was in a pleasant or unpleasant mood during recall.

of Bower's results is likely due to the fact that the subjects recorded approximately twice as many pleasant as unpleasant incidents in their diaries, and were thus drawing on a larger informational field.²⁶

In a second experiment,²⁷ subjects were hypnotized and put into a happy or sad mood, and then asked to recall as many unrelated incidents from their childhood as they could. They were then asked on the following day to categorize these incidents as pleasant or unpleasant, while in a neutral mood, in order to avoid contamination of the results by mood-contruity of interpretation. The results are shown in the next graph. A far more dramatic effect was found in the pleasant mood recall condition than in the other, but recall of non-congruous incidents was lower in both cases.

Figure 4. Number of pleasant and unpleasant childhood incidents recalled when the recaller was in a happy or sad mood.

AMERICAN PSYCHOLOGIST • FEBRUARY 1981 • 133

Still, the rather modest decrement in the sad mood condition is problematical, and in this case I can think of no obvious way to account for it. An asymmetrical dissociation such as appears to be demonstrated here, however, is sufficient for present purposes. It is also worth mentioning that we have little idea of what the actual ratios of recall are here. It might well be that the score of 10 unpleasant episodes retrieved in the sad condition represents a much higher percentage than does the score of 9 pleasant episodes, if it were the case that the subjects had had in their lives more notably happy experiences than they had outstandingly unhappy ones. It might also be that there is a general bias in favour of pleasant-event recall, and in this respect it is unfortunate

that Bower did not attempt any neutral mood trials. But in any case, these results clearly show a mood-state dependency effect -- congruent state memories are much more likely to be recalled than are non-congruent ones.

These two experiments provide evidence of mood-state-dependent dissociation in recall of significant events, and of personal history, and so make plausible the extension of the state-dependency paradigm to the everyday use of memory in cognitive judgement. We may conceptualize the effect of mood on cognition, with regard to the state dependency of episodic memory, as a shifting of the content of the field of accessible memory relative to emotional state. This is not to be construed as a total, or even predominant displacement of the informational field,²⁸ but rather a relationship such as is diagrammed below.

It would be implausible indeed to claim that little if any memory transfers between emotional states, nor would such a claim be at all justified by the experimental evidence cited. But a definite state dependency effect has been shown in retention of experimentally taught lists, and this effect has been shown to be relative to the degree of intensity of disparate moods in experiments with bipolar mania patients, and also to be relative to the degree of mood similarity (as prepared by one relevant theory) in normal subjects. Bower's evidence of state-dependency of memory for personal episodes demonstrates that this is not merely a laboratory phenomenon; there is measurable dissociation of memory when learning and attempted recall occur in different mood states. The effect of this phenomenon on the informational field drawn upon by further cognitive activities is going to be that of skewing the content of accessible memory during strong emotional states.

Before proceeding on to discussion of the relevance of mood-state-dissociation to self-deception, it will be profitable to examine some of the consequences that this sort of phenomenon has for theoretical models of human memory. Many theoretical models of long term memory, and notably the influential model of Atkinson and Shiffrin,²⁹ are based on a single framework organization of information that is not straightforwardly compatible with state dependent memory. I shall suggest that state dependent memory is most compatible with the sort of model proposed by Tulving and Thompson,³⁰ and

discuss the claim that state-dependency can be subsumed under Tulving and Thomson's encoding specificity hypothesis.³¹ There appears to be reciprocal benefit in subsuming state-dissociation as a special case of a more generally applicable theory of contextual specificity of encoding processes involved in episodic memory.

Consequences for Memory Theory

The use of information processing models derived from artificial intelligence systems to explain the function of human cognitive processes is a strategy that has become increasingly popular in the cognitive sciences, and one which has generated useful insights in many areas. The use of artificial information storage models to account by functional mimicry for the idiosyncrasies and capabilities of human long-term memory has been one of the more fruitful applications, and the most natural. The strategy here, as always, has been to supplement research and analysis with engineering.³² One attempts to design a system which will be functionally indistinguishable from human long-term memory, without structural presuppositions about what the machine is to be made of, and how it is to work. The task of developing a functional description, however, has proved to be rather a difficult struggle, unless the investigator is permitted to equip himself with structural hypotheses. These may be in the form of analogies to supposedly similar systems,

such as computers. There is nothing inherently illegitimate in the use of structural hypotheses in a functional description, no methodological imperative to keep the box labelled 'long-term memory' as black as possible. Indeed, it may be that it is only by means of structural hypotheses that a functional theory acquires intricacy and sophistication enough to advance our understanding of the phenomena, and generate testable hypotheses and points of attack.

There appears to be a conflict between state-dissociative phenomena and cognitive theories of long-term memory similar to that proposed by Shiffrin and Atkinson.³³ Shiffrin and Atkinson contrast their model of human long-term store (LTS) with computer memory, which is "location addressable" -- meaning that given the name of a storage-location the computer will deliver the contents of that location. Human LTS, say Shiffrin and Atkinson, is better understood as a "content addressable" system; which is one which, given the particular contents sought will produce a complete listing of all the locations in which those contents might be found. But more specifically, LTS is a "self-addressable" memory,³⁴ a content-addressable system in which the general parameters of what is sought contain enough information to specify the location(s) to be read out.

"This can occur if the information is originally stored in locations specified by some master plan dependent upon the contents of the information... A self-addressing memory may be compared with a library shelving system which is based upon the contents of the books."³⁵

Shiffrin and Atkinson do not claim that location is completely specified by the content of what is sought and the categories of LTS; it is, rather, that an approximate location can be demarcated in this way, which must then be searched for the appropriate word or image. But this does not alter the basic structure of this model, which involves two key suppositions:³⁶

- (i) LTS has a single organizational framework.
- (ii) Placement in LTS is fully determined by the content of the information to be stored and the organizational framework.

Nothing is said about the specific nature of this organizational framework, excepting that it must be highly complex,³⁷ and that, amongst others, it must have a temporal sorting dimension.³⁸ It also seems that functionally primitive information "bits" are to be thought of as having a single storage location, though components of these bits may, apparently, be retrieved independently.³⁹ The constitution of these "bits" is not specified, and they are referred to interchangeably as 'codes' or 'images'.⁴⁰ But the details of the model, however defensible, are not of concern here, since the challenge mounted by demonstrations of mood-state-dissociation (and, equally, of drug-state dissociation) addresses the basic form of this type of model.

According to such a model, and illustrated by the analogy to a library shelving system, the efficacy of retrieval procedures is ensured by the unidirectional symmetry of encoding and retrieval. It is because encoding procedure is determined

solely by information content and the master plan that memory can be plausibly characterized as a content-addressable system. The notable and objectionable feature of such a model is its inattention to the influence of context, and other incidental influences or encoding. Long-term memory is conceived of as an automatic system insensitive to situational context, and access to the contents of it dependent only on specificity of the information accompanying the retrieval request.⁴¹

Our review of the literature on state-dissociation has shown that memory is not uniformly accessible in all states even if it is assured that the retrieval requests are the same. Recall is demonstrably facilitated by similarity of encoding and recall states. Tulving and Thomson (1973) have demonstrated the sensitivity of long-term episodic memory to other contextual influences. In view of this disparity between content-addressable information processing models of human long-term memory and experimental evidence of de facto contextual sensitivity, it would appear to be desirable that these models be amended to account for state-dissociation and contextual effects. This is all quite in line with the objective of the top-down⁴² cybernetic/cognitive program of analysis under discussion, which is to develop functional models which accurately mimic the behaviour identified with human cognition.

The sort of modification which I wish to propose is an adjustment of the "self-addressing" feature of the Atkinson/Shiffrin model such that initial placement will be significantly determined by aspects of the encoding situation other than the

content of the information to be encoded, and such that the organizational framework of memory -- if indeed there is such a thing -- does not assure automatic access⁴³ to the approximate location of stored information on the basis of what must ultimately be semantically assessed parameters of relevance. It is significant that Shiffrin and Atkinson account for failure to access available information solely in terms of failure to localize the memory search phase sufficiently on the basis of the specification of what is to be retrieved and the restrictive power of the master framework. In the terms of their library shelving system analogy, failure to access is explained as a result of having too large a number of books amongst which to randomly search. This might result either from vagueness in the retrieval request, or as a result of an overloading of the relevant categories in the organizational framework. But the Atkinson/Shiffrin model cannot account in this way for access failure which is either systematic or intermittent despite a consistent specificity of the retrieval request. State-dissociation is an example of access failure exhibiting both of these characteristics. Some sort of modification is therefore required if the model is to accurately mimic human cognitive behavior.

I shall outline two possible approaches to the problem of amending the information processing model to enable it to accommodate state-dissociation of memory, both of which involve what I take to be an unavoidable necessity: an attenuation

"self-addressable", since both involve the abandonment of "automatic access" as a feature of the model.

One approach would be to maintain that memory is content-addressable and argue that different organizational frameworks at work in different emotional states are responsible for state-dissociation. It would not be necessary to postulate discrete frameworks, or even substantially different ones, since it would seem that fairly subtle differences would suffice. What is found to be missing at retrieval phase in a disparate state is access to the same reference system as was in effect at encoding phase.

Alternatively, one might attempt to retain the unity of the organizational framework of long-term memory in the Atkinson/Shiffrin model at the expense of the content-addressability feature. Here we would emphasize the influence of differences in perception of environmental stimuli, the influence of the already state-dissociated field of accessible past associations, and other situation variables, or the encoding and placement into the organizational framework. In this case, an adequate, retrieval request cannot be formulated in the disparate state, because the situational variables which determined the original placement to some extent are not present in the disparate state. The result is an inability to produce an appropriate retrieval strategy for lack of relevant details.

From the standpoint of our attempt to account for state dissociation, there would appear to be little to choose

between these two strategies of accommodation. It is rather difficult, in fact, to imagine what the logic of comparison would be like. Nor is there an advantage favoring the adoption of one or the other of them as a theoretical basis for our move from the evidence of state-dissociation of experimental tasks as to our view, which generalizes to the field of accessible memory as a whole. They would seem to be equally appropriate in this respect. In fact, the differences between these two systems as I see them may turn out to be rather ephemeral: the primary reason for opting for the second is that it is more directly consonant with other developments of memory theory dealing with contextual specificity of episodic memory, and can in fact be subsumed under the more general theory of encoding specificity.⁴⁴

Tulving and Thomson report results of interactions similar to those of the state dependent learning paradigm in experiments in which the context of the learning situations is systematically manipulated. A word presented as a member of a list to be learned presented in the context of another meaningful word, and later tested for recall and recognition with cues of similar and disparate contexts, was found to be far more easily recalled if similar cues were used in the recall tests. Furthermore, similar-context recall scores exceeded recognition scores in disparate cue trials.⁴⁵ in short, congruity of context was found to be a significant determinant of the effectiveness of episodic memory.

The encoding specificity principle states firstly, that specific encoding operations performed by the cognitive system at the time of encoding on the perceived stimuli determine exactly what information is stored about those stimuli. Secondly, this stored information determines the specific retrieval conditions for that information, such that retrieval is dependent on a matching of the information in the retrieval request (of which context is a large part) with the information which determined the encoding.⁴⁶ It has been suggested that the specific encoding operations referred to are mediated to some extent by input-generated associative responses to the stimuli in question.⁴⁷ This accounts for the approximation exhibited by human memory to the integration presupposed by "self-addressing" content-addressable information processing models. The important qualification here is that this is mediation, not determination. Encoding is also sensitive to situational context. In addition, the portrayal of the influence of previous memory as functioning through the medium of mediation of encoding procedures by input generated associative responses is compatible with our adopted view of long-term store as having a single organizational framework, because this decentralization of the "self-addressability" feature by means of restriction by the specificity of contexts permits the retention of the single - framework conceptualization of LTS without the insupportable corollary of universality of access which follows from any purely content-addressable single-framework system.

Subsumption of state-dissociative phenomena under the aegis of the encoding specification principle involves supposing that the encoding operations performed on a stimulus by a drugged or emotionally aroused subject are mediated by the altered state, or elements of the subject's immediate experience attributable to it, and that these are analogous to the mediation of normal cognitive operations by contextual stimuli. It also involves supposing that this mediation is responsible for the availability of retrieval information in congruent states which is not available in disparate states.⁴⁸ Translated into the framework of the encoding specificity hypothesis, the brain state present at the time of encoding can be understood cognitively as a field of information or organizational set which determines the specific encoding of stimuli presented to the system.⁴⁹ Different brain states would involve cognitive access to fields of information of different make up, due at least in part to the ongoing differences of encoding procedures. Recall failure in a disparate can be seen as the result of inappropriate or incomplete retrieval strategy,⁵⁰ in terms of Atkinson and Shiffrin's library analogy, the search is conducted in accordance with a reference strategy other than that which was used to place the book. The superiority of same state recall is the result of use of the same, or at least a highly similar reference strategy.

The library analogy isn't an especially perspicuous one for our version of the encoding specificity model, but a couple of modifications of the analogy may be helpful in at least differentiating our view from traditional information processing models.

An approximation might be made by amending the library analogy with the supposition that this library has only a subject index, and that this subject index has been compiled and is kept to date without the benefit of adequate secondary and cross-referencing. It is thereby quite possible to miss an item if one approaches it at retrieval from a different aspect than was considered primary at encoding. Now let us further suppose that this catalogue is kept by a librarian who does not consult with the library of Congress or with any other cataloguing agency, or even keep a written set of cataloguing guidelines of her own. In placing a book, she merely flips through the pages a few times, and the balance of the half-hour or so she spends on each book is devoted to the mechanics of shelving and making out the cards. As a final supposition, let us add that the holdings of this library are larger than the storage capacity of the largest and most sophisticated computer in existence.

It is difficult to see why anyone would run a library in this way, but then, as I said before, human long-term store is only very roughly analogous. Universal automatic access is rarely if ever required in cognitive activity, though it is basic to a library which must serve many different users with

different interests at the same time. It may be that an information processing system incorporating an encoding specificity principle of this type is particularly suited to the demands of an intelligent organism which may effectively depend on environmental similarities between relevant encoding and retrieval situations~~to~~ to supplement its reference systems. Such context and environment dependent delimitation of the field of stored information may facilitate retrieval of important information relevant to the retrieval situation. Fairly specific cueing, or retrieval information, is required to access stored information, but it seems likely that situations in which such information would be required would be such as to provide the necessary cues, in the normal case.

Within this theoretical framework, mood and drug-state dissociation appear as the results of wide varieties of information having been encoded under the influence of medication by state-specific associated and by mood or drug related environmental stimuli.⁵¹ The mood or drug state serves as a sort of global context within which various sorts of information are classified for placement. It is because mood and drug states involve state-specific encoding of many different sorts of information which may be generally relevant that this effect appears as a dissociation.

This subsumption of state-dissociative phenomena under the encoding specificity principle is of course extremely speculative, but advantageous, I think, in view of the

similarity of these two classes of memory-accessibility phenomena, and because of the suggestive theoretical content that the concept of encoding specificity adds to our chosen means of amending the information processing model of LTS to fit with observed state-dissociation: rejection of the context-addressability feature, retention of the single framework feature, and extension of the sense of "self-addressability" feature to include contextual and other situational variables amongst the mediators of placement at encoding phase. With these modifications, it seems that an information processing model of episodic memory consistent with state dissociative phenomena might be successfully articulated.

It is not my present intention to develop such an account any further than the suggestions that have already been offered above. The purpose of discussing models of LTS is primarily to provide a theoretical background, for the concept of state-dissociation, for the above-advocated model is not crucial to the argument that moods have important non-motivational effects on cognitive activity, which depends only on the empirical evidence of state-dissociation. The demonstrations of mood-state dependency in Bower (1981) and Weingartner (1977) are to be considered of primary importance to the explanation of self-deception.

Application to Self-Deception

The application of mood-state dissociation to self-deception involves replacement of the explanation in terms of

"motive" characteristic of the common notion and of most philosophical accounts of self-deception with an explanation in terms of "transfer failure" between disparate mood states.

It will be recalled that the defining features of self-deception, which are at least provisionally what is to be explained are apparently unjustified belief in the face of available evidence hostile to the belief, availability of that evidence to the self-deceiver (he ought to be aware of it, and of its force), such that one is tempted to ascribe a motivated avoidance of the unpleasant evidence, and the presence of emotional arousal. It is also evident that what is essentially to be explained is how it is that this sort of unjustified belief in the face of available evidence might be possible. The ascription of a motive is characteristic of common use, but we have argued earlier that it is the temptation to do this which is the defining feature, and not "motive" itself, since the ascription of motivated avoidance is in fact a kind of explanatory theory. The explanation of self-deception in terms of mood-state dissociation attributes the practice of ascribing a motive to self-deception to the attempt to make sense of a puzzling kind of unjustified belief typical of emotional states through the use of a theoretical construct.

Our account has no need of this construct. We have argued that emotional states are characteristic of the occasions of self-deception, and we have shown that mood-states cause dissociation of episodic memory. It seems reasonable to suppose that beliefs formed during, and

characteristically accompanied by, specific moods may be grounded in and supported by an incomplete field of evidence, restricted by mood-dissociation. Self-deceptive beliefs may be formed and maintained in a cognitively rational way, but this is possible because certain categories of evidence available to the self-deceiver and relevant to the belief in question are not easily accessible in the mood state associated with that belief. This will allow us to account for what intentionalists refer to as "focussing of attention" in a passive way which avoids the difficulties of intentionalist models, discussed previously. The hostile evidence away from which the intentionalists' self-deceiver must, per impossibile, focus his attention, is on our view at most merely available (stored somewhere in memory), but not easily accessible, and not in fact accessed in cases of self-deception. Self-deception is an anomaly of memory retrieval caused by mood-state dissociation.

We may illustrate this with the case of Leonard, the intelligent Reaganite with whom we made acquaintance in the first chapter. Identifying himself as a conservative nationalist and a fan of the president, we may presume that Leonard has a rather positive impression of Mr. Reagan, and approving if not exactly elated feelings associated with the president and his program. How are we to explain the fact that Leonard sees Reagan as a peacemaker, in spite of evidence to the contrary of which one would expect any reasonably intelligent person

to be aware and take into consideration? We shall assume for the sake of simplicity that Leonard is aware of this evidence and accepts its legitimacy: this sharpens the example as a case of self-deception, though it is not necessary in order to make Leonard's case one of self-deception, since a parallel argument could be developed to explain an unjustified rejection of the legitimacy of the evidence.

On our view, Leonard's unjustifiable belief is to be seen as having been formed and as being supported by normal belief formation and management policies, but based on a subset of the relevant evidence available (in memory) to him determined by mood-dissociation. His admiration of Reagan and identification with conservative nationalism are associated with moderately elated mood states to which memories of hostile evidence do not easily transfer. If we engage him in argument, we may well find that he can access that hostile evidence: he may well admit that some of Reagan's actions appear less than conciliatory to the Russians, and may even appear to modify his views while under attack. But we should not be at all surprised to hear him express the same unqualified contentions on another occasion, as if our discussion had never taken place, though if pressed he might make the same admissions as before. This is because his belief is insulated by mood-dissociation from the ravages of hostile evidence and hostile arguments alike.

Similarly, the lovesick young man who believes in the symmetry of his romance, in spite of all the evidence in the world to the contrary, holds a belief which is justified in terms of the evidence he considers, but the evidence of the depressing reality of his unrequited commitment is dissociated from his loving moods, through this evidence remains available to him, in the sense that he is not unfamiliar with it, and may well be able to remember the unsavoury incidents if given appropriate cues, or in other, less elated moods. In the loving state, however, he will remember that she admitted a rekindling of her passion for him was not impossible, but not that she insisted that the probability of this was right up there with his chances of keeping a luncheon date on Pluto.

Such is the application of mood-state dissociation to the unjustifiable beliefs characteristic of self-deception. Hostile evidence is not deliberately ignored, it rather fails to transfer to an emotional state with which it is incompatible. Nothing resembling a deliberate policy is required to account for the consistency with which the self-deceiver fails to take relevant evidence into account: this consistency is assured in our model by the association of the desirable belief and hostile evidence with incompatible mood states.

A further feature of mood-state dissociation which appears to be applicable to cases of self-deception is its defeasibility by significant cues. It will be recalled that

experiments following the state dependency paradigm which involved cued-recall testing at the recall phase did not show a state dependency effect. This result was interpreted as favouring a model of state dependency which accounted for the effect as an anomaly of retrieval, rather than encoding: if state dependency is explained as an inability to produce an appropriate memory retrieval strategy, it is easy to accommodate the absence of the effect in cued-recall experiments, because in such experiments a retrieval strategy is provided by the experimental design. Defeasibility of the dissociation at retrieval by means of appropriate cues has the following application to self-deception: we may expect that if mood-state dissociation is the mechanism responsible for the ability of the self-deceiver to hold beliefs which are totally unjustifiable in view of evidence apparently available to him (and on our account, in fact available), then it ought to be possible to induce the self-deceiver to remember the hostile evidence by means of appropriate cues. It would seem that an experiment on these lines might provide empirical verification of our thesis, but in the absence of this, it is nonetheless interesting to note the connection between this consequence of our theory and the sorts of strategies commonly adopted when we wish to "undeceive" a self-deceiver. As David Pugmire points out;

We do not inform such a person of the evidence for the ugly truth, since we accuse him precisely of being already cognizant of that evidence and persuaded of the truth. Instead we inflict on him a lingering survey of facts whose existence he admits. What dispels his false but cherished belief is not just the truth but the glaring truth.⁵²

The presentation of the evidence dissociated is of course one of the most powerful forms of cuing available where what is sought is the remembering of that information. The strategy Pugmire describes is designed to enforce recall of and attention to items of evidence of which it is believed the subject is already in some sense aware. It is also worth noting that attempts to undeceive a self-deceiver by these means would be pointless if self-deception were, again per impossibile, a deliberate ignoring of the evidence, since if such were the case even forceful reminders could be expected to be of no lasting effect.

In this chapter, we have cited evidence which suggests that mood states cause memory dissociation, made some suggestions as to how this evidence might be assimilated to memory theory, and developed an account of the sorts of unjustifiable beliefs characteristic of self-deception in terms of mood-state dissociation. We have, in this account, what I think is the most persuasive theoretical account of the referent of 'self-deception' so far proposed in the philosophical literature.

In chapter five, we will examine the practical consequences of such an account in connection with the self-deceiver's responsibility for his state, and for actions performed in that state.

Notes to Chapter Four

1. We presume that a physiological definition would also be possible.
2. Gordon Bower, "Mood and Memory", American Psychologist vol. 36, no. 2 (February 1981), pp. 129-148.
3. Distinguishing here between episodic and semantic long-term memory, following Tulving (1972 - full citation to follow), such that episodic memories are of historical events, whereas semantic memory concerns the meaning of words and applications of concepts, and their inter-relations. This distinction does not differentiate memories about events (i.e. what happened and at such and such a time) from memories of factual material (remembering something read in a book). The distinction is meant to mark out the possibility that different systems might be involved in these distinct memory functions. It might be suggested that other distinctions between types of memory could be made, but such are not required for our discussion.
4. For example, Madigan and Bollenbach (see Bower, (1981); Teasdale and Fogarty (1979); Lloyd and Lishman (1975); Henry, Weingartner, and Murphy (1973); Weingartner, Miller, and Murphy (1977)).
5. An overview of this literature is given in an article by Weingartner entitled "Human State-Dependent Learning", in Ho, Richards, and Chute (eds.), Drug Discrimination and State-Dependent Learning (1978).
6. This follows, of course, from the fact that the methodology was designed to yield informative generalizations about the effects of drugs or moods on learning and/or retention.
7. The definition of 'state of consciousness' is operational, i.e. DRUG defines one state and NO DRUG the other.
8. Weingartner, (1978).
9. Bustamante et al, "State Dependent Learning in Humans", Physiology of Behavior vol. V (1969), pp. 793-796. I have greatly simplified here.

10. For example, J.E., Eich et al, "State-Dependent Accessibility of Retrieval Cues in the Retention of a Categorized List", Journal of Verbal Learning and Behavior (1975), pp. 408-417. Another experiment, by Darley et al, Psychopharmacologia (1974) vol. 37, pp. 139-149, is flawed in its conclusion that marijuana-induced SDL is assymetrical, since the experimenters did not allow enough time between NO DRUG and DRUG trials for consolidation to take place. Hence, the transfer between states found is very suspect.
11. For example, D.A. Overton, "State Dependent Learning and Its Relevance to Alcoholism", in B. Kissen and H. Begleiter (eds.). The Biology of Alcoholism, Vol. II (1972).
12. Weingartner, Overton, and Chute express substantial agreement on this in their respective articles in Drug Discrimination and State Dependent Learning, op. cit.
13. This can be seen in the experiments of Eich et al and Bustamerte et al, previously cited. It should be noted that in cases of extremely heavy doses this feature may be masked by gross functional ablation. Moderate dosages demonstrate SDL most effectively (Weingartner (1978), p. 366).
14. Initial difficulties in replication were due to excessive drug dosages or the use of cued-recall test procedures. Eich et al demonstrated that SDL is eliminated by the use of strong retrieval cues. This feature will be discussed later in the chapter.
15. Bower (1981) used hypnotic inducement. Weingartner (1977) used manic-depressive patients and a complex mood-assessment procedure. Macht et al (1977) manipulated the moods of his subjects using electrical shock to induce anxiety. Other studies have used sleep-deprivation, pharmacologically induced depression, or clinically depressed patients.
16. Mood inducement by hypnosis is thought to be genuine, since it is supported by galvanic skin response measurements, physiological responses, and appropriate behaviour.
17. The continua in Bower's table run across the table, e.g., fear/fear = .98, fear/sadness = .69, fear/anger = .52.
18. Weingartner et al (1977), p. 288. A further discussion is to be found in Weingartner (1978). By 'cementing events together', I presume (since this was their metaphor) that what was meant was 'produces memory for sequential relation'.

19. Others are Hill et al, "State Dependent Effects of Marijuana on Human Memory" Nature 243 (1973), and Stillman et al, "State Dependent (Dissociative) Effects of Marijuana on Human Memory", Archives of General Psychiatry 31 (1974).
20. Discussed on pp. 366-67 of that volume.
21. E.g. Wickelgren, "Alcoholic Intoxication and Memory Storage Dynamics", Memory and Cognition 3, (1975).
22. Eich et al (1975), Weingartner et al (1976), and Bower (1981). This is a generally accepted finding.
23. Weingartner (1978), pp. 366-371.
24. Assymetrical SDL is a common and problematic finding, but is not regarded as detrimental to the dissociation interpretation. According to Overton (1979), models of SDL which locate the effect as an anomaly of retrival are more consistent with sometime assymetry.
25. Bower (1981), p. 133.
26. Bower (1981), p.132. Bower offers no explanation, the above is my own. Another possibility, which could not be evaluated without access to the details of Bower's procedures, is that this is some sort of ceiling effect.
27. Bower (1981), p. 133.
28. Except, perhaps, in cases of bipolar mania, and related psychological disorders.
29. Shiffrin and Atkinson, "Storage and Retrieval Processes in Long-Term Memory", Psychological Review (1969).
30. Tulving and Thomson, "Encoding Specificity and Retrieval Processes in Episodic Memory", Psychological Review (1973).
31. Tulving discusses this proposal in Tulving, "Context Effects in the Storage and Retrieval of Information in Man", Psychopharmacological Bulletin (1977), and Weingartner discusses in his (1978), pp. 377-8.
32. As in D.C. Pennett, Brainstorms.
33. I am going to pick on Shiffrin and Atkinson because their model has gained wide acceptance, and many other cognitive models share the features attached here: (i) that LTS has a single organization, and (ii) that storage location is determined only by the organizational system and the content of what is to be stored.

34. These are discussed in Shiffrin and Atkinson (1969), p. 181.
35. Shiffrin and Atkinson (1969), p. 181. Encoding and retrieval processes are conceived of as being symmetrical.
36. Ibid., p. 181, p. 184.
37. Ibid., p. 182.
38. Ibid., p. 182. No fundamental distinction between episodic and semantic memory here.
39. Ibid., p. 185.
40. Ibid., p. 183.
41. Ibid., p. 18506. Tulving and Thomson discuss this aspect of this sort of model in their (1973), p. 358.
42. For a discussion of "top-down" versus "bottom-up" analyses, see D.C. Dennett, Brainstorms.
43. Tulving and Thomson advance a similar criticism, and 'automatic access' is their term.
44. Advanced by Tulving and Thomson in a series of articles, notably (1973) previously cited.
45. Tulving and Thomson (1973). Any superiority of recall over recognition is an impossibility for a Shiffrin/Atkinson type model.
46. Tulving (1977), p. 67.
47. Tulving and Thomson (1973), p. 358-9.
48. Adapted from Tulving (1977), p. 68.
49. Adapted from Weingartner (1978), p. 377.
50. 'Strategy' in this usage refers to a set of procedures and not to an intentional plan. These procedures may well be automatic.
51. Perhaps by means of the distortion of temporal sequence discussed earlier.
52. Pugmire, p. 345.

CHAPTER FIVE

SELF-DECEPTION AND RESPONSIBILITY

The questions with which this chapter will be occupied are these:

- (i) In what sense, if any, can a self-deceiver be held to be responsible for his state of ignorance?
- (ii) In what sense, if any, can he be held responsible for acts performed in and determined to some extent by that state of ignorance?

We have characterized self-deception as a kind of breakdown of memory-retrieval between disparate mood-states. It appears that memories associated with particular mood-states tend not to be easily accessible in significantly different mood-states, and that this appears to be due to the use of context-specific cues in encoding and retrieval phases of memory, which differ from mood to mood. This does not mean, however, that mood-dissociation is something which is irremediable, or which must be accepted as being part of "human nature". Mood-dissociation is typical of human behaviour, but it is not thereby to be thought to be either inevitable or morally neutral.

Mood-specific belief, or "emotionally bound ideologies", as we might call complexes of mood-specific beliefs, are in fact familiar phenomena of human behaviour. Opinions formed, maintained, and expressed in anger, romantic reverie, and the bleak

World of depression, are important parts of the lives of many. These phenomena are related to self-deception and conducive to the development of self-deception, but they are not all cases of self-deception, and for this reason: mood dissociation is of limited duration in an integrated personality. In most cases, episodes of mood-specific belief appear as incidents of "emotional outburst" in a dominant and rational world view. In self-deception, however, the dissociation persists: what is lacking is a calm and reflective accommodation of beliefs formed in conflicting moods. We may delay for a moment the question of why this is the case, and the question of whether the self-deceiver is in an important sense free to integrate his disparate mood states: the point here is that there is an ideal of personal integration which the self-deceiver does not meet, and which he is required to meet if he is to be considered a "reasonable man". Self-deception is a characteristic of human nature, but it is a characteristic deficiency.

Nor is it an irremediable deficiency. As noted in the previous chapter, we may attempt to "undeceive" a self-deceiver by inflicting upon him a lingering survey of the obvious facts. This is not quite to the point, as it is not merely remediability but self-remediability which would seem requisite if the self-deceiver is to be held responsible for his state, but it does indicate a direction for attempts at self-remediation. Of course, one's own attempts to avoid self-deception will not be entirely analogous to the kinds

of amateur psychotherapy that one might try out on the self-deception of others. Self-remediation cannot be construed as a treatment, since the ability to self-diagnose would obviate the need for treatment. It is, rather, a question of recognizing situations in which there is a potential for self-deception and of adopting cognitive policies which will minimize one's chances of actualizing that potential. There would seem to be a number of ways of insuring by means of cognitive policy that situations in which there is a risk of self-deception will be met with a "lingering survey of the obvious facts", in order to overcome mood-dissociation: one may make an effort to reflect on emotionally disturbing issues as calmly and unaffectedly as possible so as to reduce the effect of dissociation; or one may attempt to eclectically consider the issue in all of the associated moods and integrate the result, perhaps with the aid of a journal; or one may engage the assistance of a patient listener to supplement either or both of the above with the resources of his or her presumably undissociated memory. These wholly unoriginal suggestions are no more than what people ordinarily do when trying to make rational sense out of emotionally loaded issues: this underlines the plausibility of claiming that a large part of what we are calling 'self-deception' depends on the omission of cognitive procedures for dealing with mood-dissociation which are pretty much common knowledge.

Given that self-deception is neither desirable nor irremediable, it would seem proper to attribute to the self-deceiver some responsibility for his state. The balance of this chapter will be concerned with determining the bounds of this responsibility.

As has been demonstrated in previous chapters, there is no act of self-deception, or procedure of self-deceiving. We have argued, rather, that self-deception is dependent upon a kind of negligence or omission. It is further evident that on my theory of the nature of self-deception, it is not immediately avoidable by the self-deceiver. Where the deception exists, it is already too late: the agency of the subject is subverted by the fact that he is deceived. No direct responsibility can be assessed for the state or for acts performed under the influence of this state because avoidance of it is not at this stage within the subject's power.

If the self-deceiver is to be held responsible at all, it must be for his failure to adhere to some stricter set of cognitive policies which would have prevented the development of the radical dissociation of self-deception. This sort of mediate responsibility is described by

Michael Scriven:¹

We are responsible for actions (and consequences) which we could not avoid doing at the time or avoid causing, or whose nature we did not recognize, or whose consequences we did not foresee, when there was an earlier time at which we could have acted so as to avoid the later action (or consequence, or lack of insight or foresight), at which earlier time

we did or should have known that the later result might occur and could be reasonably expected to act so as to avoid its occurrence. ... 'Should have known' is approximately equivalent to 'It was possible for him to find out, and good reasons and means for finding out were known to him, at some earlier time.'

Scriven is referring to actions, but the application to states of self-deception is fairly straightforward. We shall say that one is responsible for one's state of ignorance where the self-deception was avoidable, had strategies of cognitive policy designed to defeat mood-dissociation been undertaken at an earlier time, and for any actions performed under the influence of that state, were they foreseeable.

The general principle may be argued for as follows. Suppose that Tom sleeps in one frosty winter morning, and awakes at nine o'clock. Tom lives about a ten minute drive from downtown Ottawa, and he remembers that he has an appointment downtown at a quarter past nine. He jumps into his clothes, and rushes out to his car: after a few cursory swipes at the ice that has formed overnight on the windshield, Tom races off to his appointment. Unfortunately, however, his still ice-covered windshield prevents him from seeing an old lady who steps out from the curb: he strikes and injures her. Is Tom responsible for this accident?

He is indeed responsible, of course, even though unable to avoid hitting her at the precise moment of the accident, because there were measures which he could have taken, and of which he certainly knew, which would have greatly

reduced the chances of such an accident, i.e. scraping the windshield thoroughly. It would seem that almost all examples of negligence, criminal or otherwise, involve responsibility of this type.

In the extension of the concept of mediate responsibility to cases of self-deception, the past negligence in question is a failure to adopt appropriate strategies of cognitive management in order to overcome the effect of mood-dissociation. This failure is less blatant than Tom's failure to properly scrape his windshield, but is nonetheless within the bounds of what can be reasonably expected of a reasonable man. A certain minimum standard of personal integration is expected, and it is assumed that the attainment of this is within one's power.

The assumption that a certain minimum standard of personal integration, such as to make possible the avoidance of self-deception, is within the power of the reasonably intelligent individual requires some justification. In saying of someone that it would have been within his power for him to have avoided his self-deception, we may appear to be contradicting the thesis of universal determinism, which states that all events and conditions including those of human nature are completely determined by antecedent events and conditions. It would seem to follow from this that nothing other than what I have in fact done was ever within my power. Secondly, assuming that it has been established that there are at least

some cases in which I could have acted otherwise than I actually did, it might be that avoidance of self-deception is not one of them. And finally, assuming that a satisfactory answer can be given to the general deterministic objection, and that the claim that avoidance of self-deception is within the self-deceiver's power in most cases is acceptable, it might still seem difficult to rule out the possibility that in at least some particular cases avoidance was beyond that particular self-deceiver's intellectual means, and if so rather difficult to determine to which type of case any particular case might belong.

With regard to the general deterministic objection, a proper answer would require a demonstration that the thesis of determinism and the concept of responsibility are compatible, which is beyond the scope of this chapter. This matter has been amply discussed by others without convincing resolution. It would seem, however, that tossing the concept of responsibility out the window would be a rather drastic measure, to which a reduced sense of 'being within one's power to do otherwise' which did not involve a condition of being entirely free to do otherwise, would be preferable. From a moral point of view, it seems important to be able to distinguish between acts and conditions which are caused by factors entirely outside of an individual's control and those which involve some degree of deliberation, even if these latter are also causally determined by external factors ultimately. We shall

take it, then, that acts or conditions which are the result, at least mediately, of an individual's deliberations, or are such that deliberation might reasonably be expected of that individual, are things for which that individual may be held responsible.

Concerning the second question, which we may now frame as the question of whether the avoidance of self-deception is something at least mediately dependent upon deliberation, or a result of an absence of deliberations which might reasonably be expected, I have two points to make. First, I have argued a few pages previously that there are strategies for dealing with mood-dissociation, that these are fairly well known. Second, our social conventions regarding matters such as this support the attribution of responsibility in cases where epistemic negligence is involved: Ignorance of the law is no excuse, because it is easy enough to find out, and an obligation to do so. It seems in fact that ignorance is in general no excuse where the information in question is easily accessible and of some importance. A businessman who commits his firm to an unfavourable contract because he has not taken sufficient account of the details of the transaction will certainly be held responsible for this. Likewise, if I am misled by someone, not out of malice, but because he has not taken the trouble to inform himself sufficiently about a matter he should know of, I will want to hold him responsible for my ignorance and his own and

the consequences of both. Self-deception is perhaps most closely analogous to this last sort of epistemic negligence, excepting that in self-deception the negligent and his principal victim are the same individual. We can carry this further: supposing that I am misled by someone, not out of malice, but because he has not taken the trouble to inform himself about a matter he should know of -- and about which, in fact, we may say he is deceiving himself. Am I not, just as before, entitled to hold him responsible?

I shall take it, then, that it is in general both possible and to be expected that one should avoid self-deception, which involves the maintenance of a minimum standard of personal integration that is within one's power.

The final objection to be considered, that even if there good reasons to believe that the maintenance of a minimum standard of personal integration required for the avoidance of self-deception was generally within the power of the reasonably intelligent individual we could neither rule out individual cases in which this was not within the individual's power nor reliably distinguish this sort of case from the normal case, merits a short answer. It would seem that we ought to be able to agree that in cases of self-deception in which it clearly seems that avoidance was likely never within the power of the self-deceiver, he ought not be considered responsible for his deceived state or any consequences thereof. Such cases would be admitted as exceptions to a generally acceptable practice of holding self-deceivers

responsible. Difficulties in distinguishing these cases would not be particularly worrisome: given that the attribution of responsibility is generally appropriate as a social practice, the burden of proof will lie on the side of the cases which are claimed to be exceptional. Where such proof can be offered, we need not shrink from granting an exception to the general rule.

We can now answer the two questions with which this chapter began in the affirmative. Though self-deception is not an act or series of acts for which the self-deceiver can be held directly responsible, it results from the omission of proper cognitive procedures, the omission of which can be considered negligent. Because the avoidance of self-deception can be considered both within the power of a reasonably intelligent individual and also to be something which may reasonably be expected of him, the attribution of mediate responsibility for his state of ignorance and for any consequences following from it appears to be appropriate.

The kind of responsibility which we attribute to the self-deceiver is analogous to that which we attribute to a drunken driver. Supposing that his sluggish responses lead to an accident, we accept that at the time of the accident he could not avoid the accident, nor could he avoid having sluggish responses. It is not even to the decision to get into the car while drunk that we attach the responsibility: it is evident that this, too, may be an impaired judgement. It is nevertheless the case that many, if not most people

manage to avoid drunken driving, and that we hold those who do not avoid it responsible both for the state of drunkenness while behind the wheel and for any consequences that follow from it. I would suggest that the responsibility here, as in the case of self-deception, is a mediate responsibility attaching not to any particular act, but to one's not having done what one may be expected to have done: in this case, not having undertaken sufficient measures as to preclude drunken driving, such as training oneself never to drive after even one drink, having friends hide the keys, or whatever would be required.

In self-deception, we have a similar impairment of judgement, and a similar expectation that this is something which can and ought to be avoided by means of policies to be adopted somewhat before the fact. Mood-dissociation is the cause of insulation of certain beliefs from available evidence which contradicts them, but this condition is generally self-defeasible, and the omission of the policies required to avoid it constitutes culpable cognitive negligence.

Note to Chapter Five

1. Michael Scriven, Primary Philosophy (1966), p. 211.

BIBLIOGRAPHY

A. On Self-Deception

Canfield, J. and Gustafson, D. "Self-Deception", Analysis 23, 1962.

Demos, R. "Lying to Oneself", Journal of Philosophy 57, 1960.

de Sousa, R. "Review Discussion: Self-Deception", Inquiry 13, 1979.

de Sousa, R. "Self-Deceptive Emotions", Journal of Philosophy 75, 1978.

Drengson, A. "Critical Notice of H. Fingarette: Self-Deception", Canadian Journal of Philosophy, 1974.

Fingarette, H. Self-Deception, London: Routledge and Kegan Paul, 1969.

Frenckel-Brunswick, E. "Mechanism of Self-Deception", Journal of Social Psychology 10.

Freud, S. A General Introduction to Psychoanalysis, New York: Simon and Schuster, Inc., 1970.

Gardiner, P. "Error, Faith, and Self-Deception", Proceedings of the Aristotelian Society, vol. LXX, pp. 221-244.

King-Farlow, J. "Self-Deceiver and Sartrean Seducers", Analysis 23, 1963.

Kipp, D. "On Self-Deception", Philosophical Quarterly 30, p. 305.

Paluch, S. "Self-Deception", Inquiry 10.

Pears, D. "The Paradoxes of Self-Deception", Teorema Mono 1, 1974.

Penelhum, T. "Pleasure and Falsity", American Philosophical Quarterly, vol. 1, no. 2, 1964.

Poff-Clifton, D. Self-Deception, Carleton University M.A. Thesis, 1979.

Pugmire, D. "'Strong' Self-Deception", Inquiry 12,

Rorty, A.O. "Belief and Self-Deception", Inquiry 15, pp. 387-410.

Sartre, J.P. Being and Nothingness, Washington Square Press: 1966.

Siegler, F.A. "Demos on Lying to Oneself", Journal of Philosophy 59, 1962.

Siegler, F.A. "An Analysis of Self-Deception", Noûs 2, 1968.

Szabados, B. "Fingarette on Self-Deception", Philosophical Papers 6.

Szabados, B. "Morality of Self-Deception", Dialogue (Can.) 13.

Szabados, B. "Self-Deception", Canadian Journal of Philosophy 4, Sept. 1974.

Szabados, B. "Wishful Thinking and Self-Deception", Analysis 33.

Wilson, C. "Self-Deception and Psychological Realism", Philosophical Investigations 3.

B. On State-Dependency and Memory Theory

Bower, G. "Mood and Memory", American Psychologist, vol. 36, no. 2, (Feb. 1981), pp. 129-148.

Bustamente et al, "State-Dependent Learning in Humans", Physiology of Behavior 5, 1979, pp. 793-796.

Darley et al, "The Nature of Storage Deficits and State-Dependent Retention under Marijuana", Psychopharmacologia 37, 1974, pp. 139-149.

Eich, J.E. et al, "State-Dependent Accessibility of Retrieval Cues in the Retention of a Categorized List", Journal of Verbal Learning and Behaviour, 1975, pp. 408-417.

Henry, G., Weingartner, H., and Murphy, D.L., "Influence of Affective States and Psycho-Active Drugs on Verbal Learning and Behavior", American Journal of Psychiatry 130, 1973, pp. 966-971.

Hill et al, "State-Dependent Effects of Marijuana on Human Memory", Nature 243, 1973.

Ho, B.T. et al (eds.), Drug Discrimination and State-Dependent Learning, New York: Academic Press, 1978.

Lloyd, G.G. and Lishman, W.A. "Effect of Depression on the Speed of Recall of Pleasant and Unpleasant Experiences", Psychological Medicine 5, 1975, pp. 173-180.

Macht et al, "The Effects of Manipulations of Affective State on Free Recall and Performance", paper at a meeting of the Eastern Psychological Association, New York, 1976.

Madigan, R.J., and Bollenbach, A.K. "The Effects of Induced Moods on the Retrieval of Episodic and Semantic Memories", unpublished manuscript, University of Alaska, Department of Psychology, 1980.

Overton, D.A., "State-Dependent Learning and its Relevance to Alcoholism", in Kissen and Begleiter (eds.), The Biology of Alcoholism, vol. 2, 1972.

Shiffrin and Atkinson, "Storage and Retrieval Processes in Long-Term Memory", Psychological Review, 1969.

Stillman et al, "State-Dependent (Dissociative) Effects of Marijuana on Human Memory", Archives of General Psychiatry 31, 1974.

Teasdale, J.D., and Fogarty, F.J., "Differential Effects of Induced Mood on Retrieval of Pleasant and Unpleasant Events from Episodic Memory", Journal of Abnormal Psychology 88, 1979, pp. 248-257.

Tulving, E., "Context Effects in the Storage and Retrieval of Information in Man", Psychopharmacological Bulletin 86, 1977.

Tulving and Thomson, "Encoding Specificity and Retrieval Processes in Episodic Memory", Psychological Review, 1973.

Weingartner, H. and Murphy, D.L., "Brain States and Memory: State Dependent Storage and Retrieval of Information", Psychopharmacological Bulletin 86, 1977.

Weingartner, H., Miller, H. and Murphy, D.L., "Mood-State-Dependent Retrieval of Verbal Associations", Journal of Abnormal Psychology 86, 1977, pp. 276-284.

Weingartner, H., "Human State-Dependent Learning", in Ho et al (eds.), Drug Discrimination and State-Dependent Learning, (1978).

Wickelgren, W.A., "Alcoholic Intoxication and Memory Storage Dynamics", Memory and Cognition 3, 1975, pp. 385-389.

C. General

Churchland, Pat. "A Perspective on Mind-Brain Research",
Journal of Philosophy 78.

Dennett, D.C., Brainstorms, Bradford Books: 1978.

de Sousa, R., "The Rationality of Emotions", Dialogue (Can.) 18.

Putnan, Hilary, "The Meaning of 'Meaning'", in Mind, Language, and Reality, Philosophical Papers vol. 2, Cambridge University Press: 1975.

Rorty, A.O., "Explaining Emotions", Journal of Philosophy 75, 1978.

Scriven, M., Primary Philosophy, McGraw-Hill: 1966.

END

1041019H84

FIN