# A Defense of a Deflationary Theory of Self-Deception

by

**Kevin Lynch**

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Philosophy

University of Warwick, Department of Philosophy

July 2011

# Contents

# Acknowledgments

I would like to thank very much my parents for their support, including their generous financial support, for my writing this thesis. Thanks are also due to the Department of Philosophy at Warwick who gave me a generous scholarship to support this work. I would also like to thank my girlfriend, Lucy Cai, for her help and encouragement. The content of this thesis has benefited from discussions and feedback with a number of people. First and foremost I would like to thank my supervisor Johannes Roessler, for his intensive and steadfast support, and intellectual input, which led to great improvements in the thesis. Other members of staff have also influenced this thesis positively from discussions, including Matt Soteriou, Guy Longworth, and Naomi Eilan. I also received helpful feedback from audiences at conferences and talks on material from this thesis which I benefited from, including at the universities of Manchester, Tokyo, Bochum, Southern University of Denmark, and especially from at Warwick University where I presented material from this thesis a number of times.

# Declaration

Some material from this thesis has appeared in published papers of mine. Material from chapter 3 and 5 has appeared in an article in *Abstracta*, vol.5, called 'Prospects for an Intentionalist Theory of Self-Deception', and some from chapter 3 is in a comment piece in *The Heythrop Journal*, vol.51, called 'Self-Deception, Religious Belief, and the False Belief Condition'. An article based on the final chapter is also forthcoming in *Philosophical Psychology*, called 'On the "Tension" Inherent in Self-Deception'.

# Summary

In this thesis I take the basic idea of self-deception to be that of believing something against good evidence to the contrary because you want it to be true. I then identify the central theoretical problem concerning this phenomenon as being that of giving an account of the explanatory connection between the desire and belief in real life cases of this sort. The two main approaches to answering this question in the philosophical literature are traditionalism and deflationism. Traditionalists hold that the desire leads to the belief by motivating the subject to intentionally acquire the belief, a belief he/she initially knows to be false/unwarranted (i.e., it motivates her to intentionally deceive herself), while deflationists deny this.

I defend a deflationary account of how self-deceivers end up with their unwarranted beliefs, but one which differs from other deflationist accounts, either in substance or in emphasis, by trying to preserve a central role for agency and intentional action in the explanation and by trying to elucidate the nature of these culpable actions and intentions. Accordingly, an account is developed which holds that self-deceivers end up self-deceived because of their own actions, actions motivated by the relevant desires and emotions, though these actions are not done with an intention to deceive oneself. I try to show how an account of this sort can explain features of self-deception which any such account would be expected to explain, and in a better way than its rivals, including, for instance, the tension of self-deception, and the fact that self-deceivers are responsible for their own self-deception, while also avoiding the paradoxes that afflict other agency-focused approaches.

# Preface

The task of understanding self-deception is one which quickly embroils the investigator in questions of both a philosophical and psychological nature. Classical philosophical questions call out for answers. These include questions of conceptual analysis, where one tries to establish whether certain conditions are necessary, or sufficient, for the phenomenon to obtain. Efforts are made to find those features which distinguish the phenomenon from other related phenomenon, thus sharpening our view of how it fits into the conceptual or psychological landscape. 'How possible' questions frequently arise, where one seeks to conceive of how, if at all, it is possible for something to be the case which seems, on first look, to be impossible, and attempts are made to resolve certain logical paradoxes, or to identify fallacious presuppositions or patterns of reasoning giving rise to them. Also, psychological questions quickly arise, in particular, questions of psychological explanation, of explaining, for instance, how desires and emotions tend to cause bias in our attitudes. Furthermore, certain theories held by philosophers also frequently contain empirical assumptions which it is the job of psychology to evaluate.

Earlier philosophical work on self-deception tended not to engage much with the psychological literature, and although much progress was made in resolving some philosophical issues, there may have been a deficit present there in our understanding of the psychological processes of self-deception, which is an important aspect of the phenomenon. For sure, the keen psychological perceptiveness of certain thinkers often produced insights into the mind and behaviour of the self-deceiver, but armchair psychology could also lead philosophers in wayward and speculative directions.

Similar criticisms of the psychological literatures' avoidance of sustained philosophical discussion could also be made. One of the most striking things about the psychological literature on self-deception is how the term 'self-deception' is conspicuously

absent from so much of it. There may be a number of reasons for this, one of which may be a reluctance to get sidetracked in vexing questions of definition. Instead, these psychologists operate with invented technical terms for the phenomena they wish to study, such as 'motivated cognition'. Nevertheless, there comes a stage where we wish to know precisely what phenomenon is under investigation here, and how these studies relate to the phenomena spoken of with our everyday psychological conceptual scheme. For this the philosophical questions can't be ignored. At other times, when psychologists do explicitly present themselves as being students of self-deception, they have adopted definitions of self-deception from the philosophical literature which are in fact highly controversial, and have let them guide their empirical investigations, for better or for worse (e.g. Gur and Sackeim 1979).

Perhaps most would agree nowadays that both the philosophical and psychological work on self-deception can be mutually enriching, and interdisciplinary work on it is now a more familiar sight. I wish to continue this trend in the process of trying to answer the questions of what self-deception is and how it happens, the former being more to do with philosophy and the latter more to do with psychology.

The issue of self-deception has begat a substantial literature, and as one would hope from that fact, much progress has been made towards reaching an understanding of the phenomenon. Thus the understanding given in this thesis is not intended to be a radical departure from those previous given. I do, however, intend to provide a reasonably comprehensive treatment of the issues and to collect and build on the best insights made into the phenomena, and also to extend our understanding in new directions. I will argue that a deflationist theory is better than the traditionalist one for accounting for and making sense of the phenomena which we call self-deception, though this deflationist theory will differ in a

number of ways from previous ones, and will preserve important elements found in traditionalist views.

In the 1st chapter, after taking self-deception to involve believing unwarrantedly that *p* because one desires that *p*, I claim that the main theoretical question about self-deception is that of explaining the nature of the explanatory connection between the desire and belief in real life cases of this sort. I then distinguish between two very different types of answer we may give to this question, a 'motivational' one, where the desire motivates intentional actions which lead to the belief, or a 'merely causal' kind of answer, where the desire causes the belief via some non-rational processes. I show that with reference to the two main approaches to answering this explanatory question, intentionalism (part of the 'traditionalist' view of self-deception) and non-intentionalism (part of the 'deflationist' view), a number of philosophers have assumed that non-intentionalism is committed to the merely causal kind of answer. This is a mistake, and I emphasize the possibility of a motivationalist non-intentionalist theory, which is the sort I wish to develop in this thesis.

In chapter 2, I discuss some empirical evidence which shows the kind of behaviour involved in cases where people hold unwarranted beliefs because of a desire. I then offer a non-intentionalist interpretation of the intention these actions are done with, alternative from the intentionalist one of seeing it as an intention to acquire the belief.

Some philosophers would argue that the phenomenon discussed previously with my non-intentionalist interpretation of it, would not count as self-deception because it's a logically necessary condition of self-deception that self-deceivers intentionally acquire their unwarranted belief (i.e. intentionally deceive themselves). In the 3rd chapter I argue against this conceptual claim and defend the point that this phenomenon would count as self-deception.

Then, in the 4th chapter, I address an argument which claims that the thesis that self-deceivers intentionally deceive themselves in such cases is *explanatorily* necessary (i.e. a hypothesis that's necessary to explain how such cases come about), and I argue that this is not the case.

In chapter 5, I begin a more offensive approach to intentionalism. As I argue, doing something intentionally entails doing it knowingly, and from this we can conclude that under the assumption that the intentionalist interpretation of the intentions in these cases holds, these cases couldn't possibly have obtained, and that by a process of elimination we must therefore endorse the non-intentionalist interpretation of these cases offered in chapter 2.

In chapter 6, I discuss a particular kind of behaviour which has been said to generate self-deception by intentionalists, different from the kind of behaviour in evidence in the empirical studies we looked at in chapter 2. This strategy involving the control of attention is, I argue, not susceptible to the criticisms of chapter 5, because of the fact that it is (supposed) to be a knowledge undermining strategy. However, as I show, the idea that this strategy could work is opposed by a certain body of empirical evidence: the thought-suppression literature, and so intentionalists can find no refuge in this version of the theory.

For the final two chapters I turn to the question of whether self-deceivers have contradictory beliefs as stated in 'traditionalist' theories. In chapter 7 I discuss how it might be possible for a subject to have contradictory beliefs and argue that it would only be possible if one of them was 'unconscious', though I claim that the empirical evidence looked at in chapter 6 shows that people don't have the power to intentionally make their beliefs unconscious, and so self-deceivers don't have contradictory beliefs.

In the eighth and final chapter, I address the objection to deflationary/non-intentionalist accounts of the sort I advocate that without supposing that the self-deceiver has contradictory beliefs, we won't be able to explain the 'tension' associated with self-

deception. I argue that we can satisfactorily explain this tension from within a deflationist framework if we think of self-deceivers as having an unwarranted degree of conviction in the relevant proposition instead of an unwarranted belief regarding it, and this alternative way of conceiving of the attitude of the self-deceiver is developed in detail

## Chapter 1: Introduction

*1. The Basic Scenario and Self-Deception.*

T.R. Sarbin (1981) mentions a study done in 1969 under the direction of Edward E. Opton. Intensive interviews of 42 Americans were conducted concerning recently published photographs and stories of the My Lai Massacre of unarmed civilians by U.S. soldiers during the Vietnam War. The data was supplemented by surveys done by various newspapers and magazines. 'About two-thirds of the respondents expressed attitudes that Opton epitomized in the title of his report: "It Didn't Happen and Besides They Deserved It"' (1981: 221). Interviewees were shown the photographs and were encouraged to comment. The following were typical of the majority of responses:

> I don't believe it happened. The story was planted by Viet Cong sympathizers and people inside the country who are trying to get us out of Vietnam.

> I can't believe that a massacre was committed by our boys. It's contrary to everything I've learned about America.

> I can't believe anyone from this country would do that sort of thing.

Sarbin speaks of the 'refusal to believe that American soldiers would engage in atrocities in the face of the type of evidence that is ordinarily granted credibility (newspaper and news magazine stories and photographs)' (1981: 221).

Presumably these subjects were very averse to it being true that their own countrymen, who were in a foreign land fighting somewhat in their name, had committed

such heinous acts. Under that presumption, we suspect that their desire that this not be true had a hand in causing them to believe it wasn't true, and that a more impartial judge of the same evidence would have concluded that this event did, in all probability, occur. Given that, these cases are illustrative of phenomena possessing the following features:

A) *S* encounters evidence warranting the belief that not-*p*.

B) *S* strongly desires that *p*.

C) Because *S* desires that *p*, *S* ends up believing that *p*.[1]

Let me call these features taken together, the *Basic Scenario*. This phenomenon (in short, that of *believing something against significant evidence to the contrary because you want it to be true*), in light of the above example, does seem to be something that actually occurs.

In a moment I will discuss how the Basic Scenario relates to self-deception, but first let's acknowledge the question that immediately hits us once we note the existence of this. I'll call it the *explanatory question*. This question is:

> How, in ordinary Basic Scenario cases, does *S* end up believing that *p* because of his/her desire that *p*?

That is, the subject believes that *p* against good evidence to the contrary, where he would have concluded that not-*p* were it not for his desiring that *p*. So the question suggests itself: how did that desire lead the subject to believe that *p* against the thrust of the evidence,

[1] Note that for the remainder of this thesis, I will consistently use '*p*' to represent the proposition that is desirable for *S*, and 'not-*p*' to represent the proposition that's undesirable for *S* (except when talking about 'twisted self-deception'). I cannot guarantee, of course, that this will be the case where I am quoting others who also use these variables.

evidence which he was otherwise competent to assess? What's the nature of the explanatory connection between the desire and belief in real-life cases of that sort?

And why do we want to know this? Because it's *prima facie very puzzling* how such a thing could happen! It's puzzling, because when explaining why people believe things we typically refer to the evidence/grounds they had, or thought they had, for the belief, or mention facts which obliquely indicate what the evidence might be (e.g., he knows (believes) Gdansk is in Poland because his wife's from there). But her desiring that *p obviously* doesn't give *S* any evidence *at all* for believing that *p*, *much less* evidence that could overturn the significant evidence for not-*p*, and *S* could hardly be mistaking it for such. So how on Earth, in that case, did she end up believing that *p* from desiring that *p*?

What is the relation between the Basic Scenario and self-deception then? Let me first say that nobody would claim that features A-C are jointly *sufficient* for self-deception. However, I gather that a lot of philosophers would agree that self-deception, or paradigmatic cases thereof[2], involves *at least* features A-C[3], but whether any such case counts as self-deception will largely depend on the nature of the explanatory connection between the desire and belief. For one could think up of all sorts of such cases with 'deviant causal chains' between the desire and belief which we would have no inclination to regard as self-deception (desire triggers brain-implant to cause belief, etc.). Surely a connection of a *certain sort* would be required to make a Basic Scenario case one of self-deception (and views then vary as to the sort of connection required). At least this, I believe, would be the view of those philosophers who hold that self-deception involves unwarranted belief, which is the majority view. Others dispute this basic point, holding that self-deceivers typically don't believe the unwarranted proposition that *p*, with some saying instead that they know the truth and only pretend that *p* (Gendler 2007, Martin 1979). I will simply assume the unwarranted belief view

[2] That is, 'straight' rather than 'twisted' cases. The distinction will be discussed shortly.

[3] See the opening remarks in Boyd 2006, and also, Michel and Newen 2010: 731-732.

for the meanwhile, though in the final chapter, I will defend it against views like Gendler's and Martin's.

We should note, however, that Basic Scenario cases with deviant chains tend to be fantastical, and many philosophers expect that in *standard, ordinary cases* satisfying A-C, exemplified perhaps by the My Lai Massacre case mentioned above, the explanatory connection between the desire and belief will be of the sort required for self-deception. So they assume that actual, commonplace Basic Scenario cases are self-deception, and that answering the explanatory question for those cases will tell us what the *process of self-deception* is, i.e., the psychological process by which people end up self-deceived. This is an assumption I will make also, and it is one which, I believe, will be vindicated by the results of the investigation, in that the processes we will find in common Basic Scenario cases will be exactly what one would expect to find in self-deception. With this promise of future vindication, I ask the reader to tentatively make the assumption with me that ordinary Basic Scenario cases represent paradigmatic self-deception.

When I say 'paradigmatic' self-deception, however, I mean to contrast this with forms of self-deception which would not strike us as paradigmatic. What in the literature is often called 'twisted' self-deception would be regarded as a non-paradigmatic form of self-deception, for instance (and I don't want to suggest it's the only non-paradigmatic form of self-deception). This is where one has a fear that or strong aversion towards *p*, rather than a desire that *p*, and yet one ends up unwarrantedly believing that *p*, because, somehow, of that fear. An excellent example of so-called twisted self-deception is in the movie *Raging Bull*, where the main character Jake LaMotta's (played by Robert DeNiro) exorbitant sexual jealousy and suspiciousness towards his wife destroys his relationships, culminating in a scene where he badly beats up his brother in a rage because of his false and unwarranted belief that he slept with her, a conclusion he jumped to on the flimsiest of grounds. In the

literature these cases are generally treated as a special species of self-deception. I must say that I have never felt completely comfortable with this application of the expression. In my experience, I have never heard the expression 'self-deception' being used for such cases outside of philosophy, and I have always found it to have an unnatural ring. What I have heard with respect to such cases is the expression 'to convince oneself'. It seems much more natural to speak of LaMotta as having *convinced himself* that his brother was philandering with his wife. Colloquial usage may suggest that deceiving oneself and convincing oneself are to be distinguished somewhat along the lines on which straight and twisted self-deception have been distinguished in the literature.[4] However, I will not insist upon this point here, but will just say that whatever these cases count as, and interesting and related as they are, we will not be investigating them here. The philosophical and explanatory problems that I wish to deal with emerge specifically from the debates over the more familiar 'straight' cases of self-deception, and though these 'twisted' cases could not be so casually set aside if our goal was one of ascertaining a set of necessary and sufficient conditions for self-deception, this is not a challenge I am taking up in this thesis. So if the reader tends to regards cases of this sort as a species of self-deception, he/she may take the concern of this thesis to be restricted to coming to an understanding of *straight* self-deception, where the belief that one is self-deceived in holding is welcome.[5]

So again, in seeking an answer to the explanatory question we are trying to establish the process of (straight/paradigmatic) self-deception, and this will be the main concern of this

[4] Alternatively, there may be some indeterminacy in the notion of self-deception regarding this matter. Intuitions on what to say may vary or we may not know what to say. Experimental work on what peoples' intuitions actually are on these questions could be helpful here.

[5] For a sophisticated philosophical study of 'twisted self-deception', see (Barnes 1997). Barnes defends an analysis of self-deception according to which it includes cases where the belief one is self-deceived in holding may be either welcome or unwelcome (both 'straight' and 'twisted' cases), and she gives a unified explanation of these beliefs as existing for the purpose of reducing anxiety, though in unwelcome belief cases, the anxiety is grounded in a desire that is not about *p*, but about another proposition, *q*.

thesis, particularly in part I. In seeking this answer, we will be particularly interested in establishing whether the right explanation is of a *non-rational* kind, where, say, the desire triggers some involuntary causal process or causally influences some aspect of cognition, thus leading to the belief, or whether the explanation is of a *rationalizing* kind, where the desire motivates voluntary, intentional actions which lead to the belief. And if the latter, we will want to know what the description of the intention is here, and in particular, whether it is an intention to acquire that belief. These issues are deeply relevant to the understanding of self-deception and to some important disputes surrounding that phenomenon.

The two main and competing philosophical approaches to understanding self-deception are those of *traditionalism*, and *deflationism*. And accordingly, we can understand the primary difference between the two main theoretical camps to lie in their provision of different answers to the explanatory question in typical Basic Scenario cases (which they take to represent paradigmatic self-deception), with the former providing an answer in terms of intentional efforts on the self-deceiver's part to induce that belief in herself, and the latter rejecting this idea. The term 'traditionalism' is often used to denote theoretical persuasions that dominated in earlier work on the topic. Early proponents of this kind of perspective who were most read include Pears, Fingarette, Demos, and Davidson. The deflationist approach to self-deception developed in critical response to traditionalism, and repudiates some of its core assumptions. Alfred Mele, Annette Barnes, and Ariela Lazar are noteworthy proponents of this view.

Before we look at these theoretical approaches in any more detail however, it first will be necessary to outline two fundamentally different forms that an account of the explanatory connection between the desire and the belief may take, which I will call a 'merely causal' form and a 'motivational' form (which could also be called a non-rationalizing and rationalizing form). Then we may position traditionalism and deflationism with respect to

these. This will also give us the opportunity to define the term 'motivation' more precisely, an important term in the self-deception literature but one which is often used in too unconstrained a fashion. I see this as being an important preliminary because it seems to me that there is a mistaken view present in the literature over what the commitments of deflationism are, where it's assumed that deflationism is committed to the idea that the desire explains the belief in a 'merely causal' and *not* a 'motivational' way (in my sense of the term). This mistake has had significant negative consequences for the dialectic about self-deception: narrowing our view of the theoretical options available to us in this debate, and obscuring one attractive option in particular which I wish to develop in this thesis.

## *2. Merely Causal and Motivational Influence.*

There are two distinct explanatory roles which desires (that is, the condition of desiring something[6]), can take which will be important for categorizing answers to the explanatory question, first, what philosophers often called a 'rationalizing' role, and second, a non-rationalizing or what I'll call a 'merely causal' role. I wish to avoid getting bogged down in the difficult task of elucidating this distinction in general terms, preferring to do it by illustrating with examples. However, one might say that with respect to the former, I mean to refer to the role a desire may play in rationally explaining intentional, intelligent actions (and omissions). This is where we say that the subject did something (partly) because he had that desire, where the fact that he had that desire makes or helps to make rational sense of the

[6] The expression 'a desire' has, as Maria Alvarez remarks an 'act/object ambiguity': it can 'refer to someone's desiring something, on the one hand, and to what is desired, on the other. For example, 'my desire to go on holiday' is sometimes used to refer to *my desiring* to go on holiday, and sometimes to *what* I desire, namely to go on holiday (2010: 65). Here, by 'a desire' I mean the former, i.e. the psychological condition of desiring something.

action, helping us to see what the attraction was in doing it from the agent's perspective.[7] However, the fact that a person desired something can also be cited in a very different kind of explanation, where it explains some bodily or psychological change or event with the subject, which will in this case not be an intentional action. In these cases, the desire may stand to the explanandum as non-rational cause to an effect. The following will hopefully illustrate this point.

The distinction may apply to the case of fear. For instance, we would be dealing with rational explanation if we explained your avoidance of a certain isolated alley on the way home by mentioning your fear of/aversion to getting mugged or attacked. The explanandum here is an intentional action or decision concerning how to act which your fear helps 'make rational sense' of, generally in combination with the assumption that you had certain beliefs (e.g. that muggings sometimes take place in dark alleys). Looking for an example of 'merely causal' explanation, there is a difficulty in drawing the line between things that might be an effect of your fear, and things that are partly constitutive of it, if indeed there is such a sharp line to draw (is one's increased heart-rate, or the pang one might get in one's stomach, an effect, or partly constitutive of one's fear?), but an uncontroversial example might be if one suffered a heart attack from experiencing terror or fright, if one had a weak heart. Here, although the fear is explaining something, the explanandum is not an intentional action, or anything needing to be explained rationally, and here we say the emotional state caused the heart attack. With regard to desiring, this distinction may be evident in the difference between going to the shop and purchasing food because of your desire for food (an intentional action rationally explained by your desire), or your salivating at the sight of it because of your desire (an event which the desire caused but does not rationally explain).

[7] Philosophers these days typically construe this 'because' as indicative of a causal relationship, though there are some who dispute this. Though for my purposes it is not necessary for me to take a stand on this issue, the reader should feel free to conceive it like so if he/she wishes.

This is a very rough attempt at drawing the distinction, but I wish to avoid getting bogged down in attempting to elucidate it more thoroughly, since this difficult task would swiftly carry us away from our primary concerns. However, I hope it's evident that there's an important distinction to be noted here, which the reader may have an intuitive grasp of. Furthermore, for the sake of bringing order to the use of this term, I propose that 'motivation' be defined with reference to the role desires have in 'rationalizing' explanations of actions. Thus we can say that Jones' desire for food motivated his act of buying some, but not that his desire motivated his salivating, rather, it merely caused that. However, its use has not always been so confined in the literature. Mele, for instance, says that *whenever* desires enter into causal explanations, those explanations can be counted as motivational ones (2003, p.238), but this, in my view, fails to respect the essential connection between the ordinary concept of motivation and the phenomenon of agency, and it also obscures the distinction between the two importantly different explanatory roles desires (and perhaps other mental phenomena) may have.[8]

Now one important question to bear in mind when looking for an answer to the explanatory question is this: is the explanatory connection between the desire and the belief in self-deception one which involves the desire playing a motivating, or a merely causal role? Does the desire motivate intentional action, where the outcome of that action is to cause/sustain the unwarranted belief, or could it be that the desire explains the belief merely by playing the role of being a mere cause? Let's call theories advocating the former, *agentivist*, and one's advocating the latter, *non-agentivist*.

[8] Note that this use of 'motivate' may not be *completely* faithful to the ordinary use. We speak, at least more usually, of people being motivated to do things by something *non*-psychological, as when we say 'Jones was motivated by money' (i.e. by the prospect of earning money), though being motivated by money *presupposes* that one wants money. In case my use is unfaithful to the ordinary one in this way, the reader may take it to be a semi-technical use.

*Agentivism*: Self-deceivers end up with their unwarranted belief as a result of their own actions, actions motivated by the desire that *p*.

*Non-Agentivism*: Self-deceivers end up with their unwarranted belief not as a result of their own actions.

Traditional theories of self-deception have been agentivist. But to see how non-agentivism might be a plausible idea, consider that there seem to be examples of affective states influencing cognition in merely causal ways. Moods, for instance, can influence how you see the world: in a good mood you may be more likely to see the good in people, while a bad mood can engender more pessimistic and cynical attitudes. One theory of how this works is that moods exercise priming effects on 'mood-congruent' memories (that is, a good mood will increase the accessibility of associated positive memories, making them come to mind more readily, see Kunda 1999: 249-250), and this would be a purely non-rational way in which our thinking/attitudes can be influenced by an affective state. Might desires influence cognition in self-deception in similar ways?

Whether we understand the desire to cause the belief in self-deception by way of it playing a motivating or a merely causal role will have important ramifications for the issue of whether some common, pre-theoretical convictions about self-deception find vindication. For one such conviction seems to be that there *are* self-deceptive actions, that self-deception is something we *do* (Hellman, for instance, speaks of our 'intuitions that self-deception is a state determined by our behaviour and yet something that our behaviour can take us out of' (1983: 120)). This is perhaps suggested by the grammar that we use in talking about self-deception, as when we use the active 'he is deceiving himself' instead of the passive 'he is self-deceived'. On the merely causal account of the explanatory connection, however, this intuition would not be vindicated. Furthermore and relatedly, this question will have

relevance for the common assumption that a person's self-deception is something that they can be held responsible for. The idea that we end up self-deceived due to our own actions is perhaps most congenial to this assumption. On the other hand, if the unwarranted belief is more the result of the automatic influence of desire on cognition, then we would perhaps be better thought of as hapless victims of the condition, as we might be of schizophrenia. We would be responsible for our self-deception to no greater an extent than we would be for any other non-rational effect of an affective state, such as a tremor from nervousness.[9]

If the right explanation for real life cases which display features (A) to (C) turned out to be in terms of non-rational causal processes, then with respect to these ordinary assumptions we would, as I see it, have two options. First, we could give up on them, regarding them as naïve, pre-scientific theories or myths about self-deception, a phenomenon which only scientific psychology is equipped to understand. Or instead, we could hold on to them, regarding them as elements of our very concept of self-deception, such that if the correct explanation for common Basic Scenario cases were to find no place for them, then that would only show that these cases were never cases of self-deception to begin with, but of something else (wishful thinking perhaps).

My own view is that we should regard these assumptions as reflecting aspects of the concept of self-deception that should be steadfastly held on to and accommodated in any theory of self-deception, since it seems we are naturally reluctant to call something self-deception if it were not the result of the subject's own actions and if she could not reasonably

[9] I am not saying there is no sense in which a person might be blameworthy or responsible for the fact that a certain automatic effect follows from an affective state. A story might very well be told illustrating how someone can be so responsible. But the sense in which one might be responsible for the fact that certain automatic effects follow from one's affective states would not be as robust or direct as the sense in which we are responsible for our actions. It can thus surely be maintained that a theory which ascribes a motivational role to the desire in self-deception is at least more congenial to the idea that self-deceivers can be held responsible for their self-deception.

be held responsible for it. But I will not have to argue this point (insofar as a linguistic intuition could be argued for) since as will become clear from the empirical evidence I will shortly discuss, the correct explanation for real-life Basic Scenario cases *will* make reference to intentional actions.

Now that the distinction has been made between these two different kinds of explanatory roles, we can bear it in mind when considering the two main theoretical approaches of traditionalism and deflationism. In particular, my initial aim in doing this is to correct a significant mistake, sometimes voiced by deflationism's detractors, and sometimes fostered by the remarks of deflationists themselves: that this approach is committed to the idea that the desire plays a merely causal role, whereas the idea that self-deception arises from our intentional actions is the sole property of the traditionalist school of thought. This misconception has the effect of polarizing the debate, and unduly limiting our impression of the theoretical options available to us, and my main aim in this introductory chapter is to impress upon the reader the possibility of a middle position between the view that self-deceivers intentionally deceive themselves, and the position that they end up self-deceived because their desires influence their attitudes in automatic, non-rational ways. The possibility and attractiveness of such an account has not been sufficiently appreciated, and it is an approach in want of development.

### *3. Outline of the Basic Differences between Traditionalism and Deflationism.*

In order to outline the basic differences between traditionalism and deflationism, I will first lay out the core features of traditionalism, since this was the first theoretical approach to come on the scene, with deflationism developing largely in critical response to it. After having done that, deflationism should then be defined negatively in terms of the negation of

those core claims. This negative characterisation of deflationism then leaves us free to encompass with that term a number of theoretical options that have gone under that heading. Deflationism will thus be defined as any theory of self-deception which seeks to explain or analyse it without utilizing certain controversial notions that the traditionalist regards as necessary for understanding it (in the sense of them being either logically necessary, or explanatorily necessary). A more positive and specific deflationist theory will then be developed later and throughout the thesis.

The following elements, then, are posited by the traditionalist *in addition* to the Basic Scenario conditions, A to C (which may be taken as common ground between the traditionalist and deflationist), to make it a case of self-deception as the traditionalist sees it. That is, the traditionalist is already assuming that the subject in self-deception has encountered evidence warranting the belief that not-*p*, that he desires that *p*, and that this desire is involved in some causal/explanatory process determining that he believes that *p* against the thrust of this evidence. Given this situation, the traditionalist thinks that the following is what occurs in such cases:

i) On the basis of the evidence that Jones encountered–evidence which warranted the belief that not-*p*–Jones ends up warrantedly believing that not-*p*. Because of his strong desire that *p*, this realization causes him much distress.

ii) Jones' anxiety is such that he comes to believe that it would be better to believe that *p*, despite the fact that *p* may not be true, so as to avoid the anxiety associated with believing that not-*p*. Consequently he desires to believe that *p*.

iii) Jones executes a strategy with the intention of making himself believe that *p*.

iv) Due to his efforts, Jones ends up in a condition where he believes that *p* and believes that not-*p* simultaneously[10].

This can be regarded as a characterization of traditionalism in that it contains the two main features associated with traditional accounts, advanced by the likes of Davidson and Pears, which is the *intentionality feature*, given in points (i) to (iii), which very much come as a package, and the *contradictory belief feature*, (iv). The claim that self-deception is to be characterized in terms of these features is what I identify as the core claims of traditionalism.

As for deflationism, I'm defining it negatively as any approach to self-deception which tries to understand it without utilizing the two core features attributed to self-deception by traditionalists, the intentionality and the contradictory belief feature, features which have been associated with paradox and thought by some to render self-deception an impossibility (for reasons that can be discussed in detail later). If this definition of deflationism is a stipulation, then it is a principled one, for I intend for it to capture with some degree of adequacy how the term has been used and intended to be understood in the literature. By characterizing deflationism negatively here, we end up defining it in a fairly encompassing way, such as to leave open a number of possibilities as to the nature of the explanatory connection between the desire and belief.

One of the basic ambitions of the deflationist, then, is to demystify our understanding of the phenomenon by purging it of the *prima facie* paradoxical elements found in the traditionalist account, elements that largely arise from modeling our conception of self-

[10] This would be a state of what Davidson called 'internal irrationality' (2004/1982). Internal irrationality denotes any condition whereby a person has propositional attitudes which are mutually contradictory or inconsistent in an appropriately 'direct' way, such as if someone were to believe that *p* and that not-*p*, or if someone were to have an intention to do *x* and a belief that *x* is not the best thing to do all things considered. I take the standard traditionalist view here to be that the unwelcome belief is unconscious, while the welcome one is conscious. Because the welcome one is unconscious, it no longer generates the anxiety that it did before.

deception on that of interpersonal deception (more on this later). In doing so, deflationism avoids the need for postulating what Alfred Mele calls 'mental exotica', which traditionalists often end up relying on to try to explain away the paradoxicality of these notions (as with Pear's theory of 'mental subsystems', or the notion of unconscious intentions), which might appear no less puzzling, to some deflationists at least, than the paradoxes they are brought in to solve. The deflationist proceeds to develop and defend a theory of self-deception which avoids the two traditionalist features entirely, rather than endorsing them and then going to the trouble of trying to explain away their putative paradoxicality.

The relationship between points (i) to (iii) on the one hand, and point (iv) on the other, however, is not so straightforward. Indeed, some who endorse points (i) to (iii), deny that the contradictory belief feature necessarily follows from that and seem to reject this condition in their own accounts (e.g. Talbot 1995. Bermúdez 2000: 313). Others endorse the contradictory belief feature, but reject the intentionality feature (e.g. McLaughlin 1988). Such theories we may call 'semi-traditionalist', while endorsement of both these claims together may be called 'pure' traditionalism. Because of the complexities in the theoretical relation between these two conditions, and because the debate over the contradictory belief condition is not so important with respect to the explanatory question,[11] I propose that we defer discussion of the contradictory belief condition until a later stage and concentrate for the meanwhile on the debate between traditionalists and deflationists over whether the intentionality condition holds of self-deception. Philosophers who advocate this feature, and who thus think that self-deceivers intentionally deceive themselves (where this means intentionally make themselves have a belief which they knew to be false/unwarranted), are called *intentionalists*, and their

[11] The debate over the intentionality condition is a debate over what the *process* of self-deception is. The debate over the contradictory belief condition is a debate over what the resultant *condition* of self-deception is. The question of what the process of self-deception is is what is important for the explanatory question, because with this we want to know how the subject ends up with the belief because of the desire, and presumably this will involve identifying the relevant causal/explanatory processes between the desire and belief.

opponents are variously called *non/anti-intentionalists*. So with respect to this particular area of contention between traditionalism and deflationism, the traditionalist is an intentionalist, while the deflationist is a non-intentionalist, and it is the debate between intentionalism and non-intentionalism that part I of this thesis is reserved for. Part II of the thesis will be dedicated to the discussion of the contradictory belief feature, and eventually I will argue that this feature is intimately associated with the intentionalist account, or at least its most promising version.

Though our account of the difference between traditionalism and deflationism is inchoate as yet, it nevertheless puts us in a position to draw some conclusions about whether each approach regards the desire as playing a motivational or merely causal role in explaining the unwarranted belief in Basic Scenario cases.

### *4. Intentionalism and Non-Intentionalism, and the Motivating and Merely Causal Role Distinction.*

We can see that for the intentionalist, the desire most 'proximately' involved in explaining the self-deceptive action is the desire to believe that *p*, rather than the desire that *p*. The content of this desire specifies the goal of the action: to believe that *p*. On one construal of 'motivation', we might say that desires only of this sort are what motivate actions. However, this construal seems too narrow, as the desire that *p* is also a crucial protagonist in the story that makes rational sense out of what the subject is doing. This is so because Jones' reason for wanting to believe that *p* is that this will allow him to avoid the distress associated with believing that not-*p*, distress he wants to avoid. And he feels distress at this precisely because

he desires so much that *p*.[12] Thus without the assumption that Jones desired that *p*, Jones' wanting to believe that *p* wouldn't be intelligible to us. And I wish to say that when a subject does an action because of a mental state or attitude, and where that state/attitude allows us to make *rational sense* out of that action, helping to make intelligible to us the purpose or good of the action from the agents perspective, we can speak of it as 'motivating' the action.

But what about non-intentionalism? Here I wish to pre-empt the possibility of a misunderstanding about what the commitments of non-intentionalism are. For from observing that in the intentionalist account the desire that *p* explains the belief in a rationalizing way by helping to motivate intentional action, it may be tempting to think that what then differentiates non-intentionalism from this approach is that it takes the desire to explain the belief in a merely causal way. And there is sometimes an impression given in the literature that once we go non-intentionalist, we are committed to the idea that intentional action, motivated by the desire that *p*, must play no role in the explanation of the self-deceptive belief at all. For example, Talbot interprets 'anti-intentionalism' as holding that the desire causes the belief via some 'non-intentional mechanisms' that get 'triggered' or 'activated' by that desire (Talbot 1995).

On this way of distinguishing them, the theoretical divide between these two persuasions appears as particularly wide. It also makes non-intentionalism seem the intuitively less plausible option, since in making it out to differ so radically from intentionalism, non-intentionalism is made out to look antagonistic towards some key points found in intentionalism that are worth preserving. It thus puts non-intentionalism in conflict with the intuition that self-deception results from our actions. Moreover, since the

[12] This 'because' is not indicative of a contingently causal relationship. Some philosophers hold that emotions are things people can have reasons for, and perhaps we would say that the discovery that *p*, when he strongly desires that not-*p*, gives the subject *a reason* for feeling distress. Alternatively, perhaps we should look on the relation here as a logical one, rather than a 'reason giving' one, since it seems to be a conceptual truth about desire that people who strongly desire that *p* will tend to feel distressed or upset when they discover that not-*p*.

intentionalist/non-intentionalist distinction is supposed to be exhaustive of answers to the explanatory question, it hampers our ability to imagine the kind of middle-way view that I would like to develop here.

The rationalizing and merely causal role dichotomy is not mirrored in the distinction between intentionalism and non-intentionalism itself. For the supposition that the unwarranted belief that *p* forms as the result of *the person's actions* (including mental actions), actions which are *intentional*, and motivated by the desire that *p*, may be consistent with non-intentionalism, *so long* as the intentions with which these actions are done are not intentions to deceive oneself (*that* was all the non-intentionalist was concerned to deny!). Again, this point is often not appreciated. Bermúdez, for instance (like Talbot, a critic of non-intentionalism), says that '[a]ccording to anti-intentionalist accounts…self-deceiving belief formation can be explained simply in terms of motivational bias, without bringing in appeals to intentional action' (2000: 309. Also see Doucet, forthcoming. Note, incidentally, how 'motivational' is being used here in a wider sense than I use it). But though some versions of anti-intentionalism may take that form (e.g. Johnston 1988, Hales 1994), all need not. Here we ought to note that what the anti-intentionalist is 'anti' about is the idea, thought to be paradoxical, that self-deceivers intentionally deceive themselves, and this does not preclude him or her from appealing to intentional actions to answer the explanatory question if the attendant intentions fall short of being intentions to deceive. Thus the theoretical divide between intentionalism and non-intentionalism may be more subtle than is suggested in such presentations. Non-intentionalism may come in agentivist or non-agentivist varieties.

Why have some philosophers made this mistake of assuming that non-intentionalism must be non-agentive? There may be a number of reasons for this, and the attempt to specify them will be speculative to an extent. Sometimes it may simply be a case of heedlessly assuming from the look of the expression 'non/anti-intentionalism' that it means 'non/anti-

intentional action'. But another reason may be associated with the work of Alfred Mele. Mele is the most prominent and influential deflationist/non-intentionalist and people's impressions of what the view involves may derive significantly from their reading of him. It is possible that the association between non-intentionalism and merely causal explanations was picked up from reading his work, since it is natural to read him as a non-agentive deflationist. However, the question of whether Mele really holds an agentive or non-agentive theory of self-deception is not a straightforward one to answer. My own view is that on the one hand, when one looks at the details of the processes Mele cites as being operative in self-deception, processes we will look at in chapter 2, one finds a combination of rational and non-rational process, which Mele is not very concerned to distinguish and which he indiscriminately describes as 'motivational' (in one place he explicitly says that whenever desires enter into causal explanations, those explanations can be counted as motivational ones (2003: 238), and so he is using 'motivational' in a wider sense than I). But on the other hand, when Mele attempts a general characterization of his theoretical approach, he often ignores these subtleties and implies that it in non-agentive. For instance, in a number of places Mele describes the view that self-deceptive beliefs are intentionally acquired/retained, *the agency view*, and the view that they are not intentionally acquired/retained (which is his own view), *the anti-agency view* (see 1998: 353-354. 2004: 247). However, as I've stressed, one can accept the latter view and still hold that 'motivationally biased belief' occurs as a result of the subjects actions, actions rationally explained by the desire that *p*, and for that reason it is deeply misleading to describe it as *the anti-agency* view. Mele seems here to conflate the denial of the intentionalist claim, i.e. the claim that self-deceivers intentionally make themselves have their unwarranted belief, with the denial of agentivism, just as we saw with Talbot and Bermúdez.

One of the main ambitions of this thesis is to address the key question of whether a plausible agentive story can be told of the Basic Scenario, where the desire plays a motivating role in explaining the belief, but which is nevertheless non-intentionalist. I will develop a positive answer to this, thus steering a course between non-agentive non-intentionalism and intentionalism. This is desirable for a number of reasons. First, non-agentive non-intentionalism is an overreaction to the problems of intentionalism. In taking flight from the paradoxes of intentionalism, non-agentive non-intentionalists abandon certain general features present in intentionalist accounts which ought to be preserved. Furthermore, it is also desirable that we avoid the excesses of intentionalism, for as I will argue, intentionally deceiving oneself would be impossible to do, at least in ordinary circumstances, and so the supposition of intentional self-deception is not a realistic contender to explain Basic Scenario cases.

The kind of non-intentionalism I will recommend, which will be agentive, will hold that in ordinary Basic Scenario cases, which I take to be paradigmatic of self-deception, the desire that *p* be true motivates, and provides a rational explanation for, certain *intentional actions*, actions which are *biased* (though not intentionally biased), and which in virtue of that explain why the subject has a biased, unwarranted belief. However, anyone who wishes to present such a theory ought to not only give an account of what the relevant actions are, which I will do by consulting some empirical literature, but also of the intention associated with these actions, thus giving a positive account of what the self-deceiver is up to in self-deception, and showing that the intention is not an intention to deceive.

Before we begin this project of determining the nature of the self-deceptive actions and the associated intentions however, it will be very helpful to discuss in more depth the elements of the traditionalist account mentioned above to get a clearer picture of the

traditionalist position, and to show how deflationism, or my version thereof, departs from each of these points in turn.

## *5. Discussion of the Elements of the Traditionalist Account, and Deflationist Disagreement.*

Earlier, four conditions were given which help to characterize the traditionalist view of self-deception. In this section, I think it will be helpful to elaborate on them a bit more and to point out how the deflationist view I will be putting forward will differ from each in turn.

(i) 'On the basis of the evidence that Jones encountered–evidence which warranted the belief that not-*p*–Jones ends up warrantedly believing that not-*p*. Because of his strong desire that *p*, this realization causes him much distress.'

This first point concerns what first happens after Jones encounters the unwelcome evidence, and this gives us the opportunity to see the first important difference between the traditionalist and deflationist approaches. For the traditionalist, we said that the subject starts out drawing the warranted conclusion on the basis of the evidence. The deflationist, on the other hand, will deny that the warranted conclusion gets drawn at all from the unwelcome evidence, though the subject will recognize that the disagreeable proposition is suggested by it, and may feel worried that it might be true that not-*p*. For the deflationist, the self-deceiver, after encountering this evidence*, assesses* it in a manner that is biased by his preferences from the very start, and this skewed assessment doesn't allow a warranted attitude to get hold in the first place. However, on the traditionalist view, the self-deceptive activity happens *after*, *not* during, the assessment of the data. So to put it another way, for the intentionalist, the data

gets assessed in a relatively accurate way to begin with, and this causes the alarm that leads to the reasoning of (ii) and the intentional action of (iii), while for the non-intentionalist, the data never gets assessed in a relatively accurate way to begin with; the subject's evaluation of the evidence is biased from the word go. Later it should become clearer what is motivating the non-intentionalist to deny that the self-deceiver ever draws the warranted conclusion to begin with.

Note that there may be 'weaker' traditionalist cases possible where the subject doesn't end up believing that not-*p*. For instance, he might come to believe that not-*p* is *probably* true, or might come to have a strong *suspicion* that not-*p*. What is important for this view is that the subject come to some settled, unwelcome attitude which is sufficient to cause the alarm that would induce the subsequent reasoning. For simplicity's sake I will stick with the more commonly discussed case where the subject simply warrantedly comes to believe that not-*p*. Because the difference between the stronger and weaker cases is only a matter of degree, doing so should not cause any problems. Any theoretical and philosophical difficulties that there may be associated with the stronger cases should be applicable to the weaker cases and vice versa, since there are no fundamental, qualitative differences between them.

(ii): 'Jones' anxiety is such that he reasons/judges that it would be better to believe that *p*, despite the fact that *p* may not be true, so as to avoid the anxiety associated with believing that not-*p*. Consequently he desires to believe that *p*.'

Point (i) stated that the proposition, not-*p*, which Jones initially believes on the basis of *C*, is unwelcome. So Jones desires that *p*. But with point (ii), we see that on this account, the self-deceiver also desires *to believe* that *p*, and this marks another point of difference with the deflationist, who denies the subject has such a desire. We should be clear about the sense

in which this is intended. In ordinary conversation, we might say things like 'I would like to believe that we are going to have a good summer this year, but you never know with this country's weather'. This would just be a way of saying that you would like for this summer to be a pleasant one. The concern is not directly with our own psychological condition, despite how that statement looks on the surface. If we were to offer someone who said this a drug, telling him that it would solve his problem and induce the belief that we'll have a pleasant summer, but would do nothing to make that belief true, he would surely tell us that we have misunderstood him. It is not his own state of *belief* he is concerned about, rather, it is the weather. He is not expressing a wish to be deluded, indeed, the idea would probably sound perverse to him.

However, the intentionalist thinks that Jones literally wants to believe that *p regardless of, or in spite of, the fact that he knows it's false/unwarranted*. To use a somewhat crude expression, we might say that the focus of his concern is now 'inwardly directed', that is, it is on changing his *psychological* condition. For the deflationist, however, the subject does not have this desire to believe that *p*. For one, the deflationist believes that the self-deceptive belief can be explained with reference to the subject's desire that *p*, without needing to refer to any desire to believe that *p*. And the other important reason why the desire to believe that *p* is not invoked by the deflationist is that this idea is too closely associated with the idea that self-deceivers intentionally deceive themselves (i.e. act with the intention of *making themselves believe* something), which deflationists are very concerned to reject. For if one was to *act* on this desire *to believe* that *p*, that is, if one was to try to satisfy this desire, then this would amount to one acting with the intention to make oneself believe something (which one knows is unwarranted), which amounts to an intention to deceive oneself.

(iii): ‘Jones executes a strategy with the intention of making himself believe that *p*.’

For this point I would like to make clear the practical reasoning associated with the execution of this strategy, and relatedly, the element of rationality that the intentionalist associates with this phenomenon. Here we see that for the traditionalist, who with respect to this point is an intentionalist, Jones acts with the intention of satisfying his previously mentioned desire to believe that *p*. To accomplish this, he executes what is called a *strategy* of self-deception. What these strategies might involve need not concern us for now, though it is an important issue which will be discussed later. Given that Jones thus intends to make himself acquire a belief which he also believes to be false or at least unlikely and unwarranted, it is natural to describe this intention as an intention to deceive himself (after all, if one person intends to make *another* person believe something which she takes to be false/unwarranted, this would ordinarily be called an intention on her part to deceive that person). We can now see that, on this account, the self-induced condition of self-deception is going to be the outcome of a piece of *practical reasoning* on the part of the Jones. Davidson, for one, is quite explicit about this. About the self-deceiver he writes that, ‘[h]is practical reasoning is straightforward. Other things being equal, it is better to avoid pain; believing [that not-*p*] is painful; therefore (other things being equal) it is better to avoid believing [that not-*p*]’ (2004/1986: 209). If we were to try to construct the practical syllogism associated with the intentional action, we might do so as follows:

- It would be desirable to acquire the belief that *p*.
- By doing *V*, I can best acquire the belief that *p*.
- I should do *V*.

Here, *V* stands for the strategy of self-deception, whatever this may be. By presenting self-deception as the outcome of practical reasoning, the intentionalist tries to get us to see an element of rationality behind this ostensibly irrational phenomenon. That is, the intentionalist sees the person who deceives himself as acting in the rational pursuit of what he regards as his own best interest, and the phenomenon is regarded as an example of attempted 'utility maximization' (see Talbot 1995). Here it is common to find proponents distinguishing between different senses of rationality, *theoretical* or *epistemic* rationality, and *practical* or *prudential* rationality (again, see Talbot 1995: 55). Though the subject may see the belief as being 'irrational' to hold in the sense of it being unwarranted by the evidence (i.e. it's 'epistemically irrational'), he may see it as being rational or best to acquire it anyway in the sense that this may minimize his total experience of anxiety (which would make it 'prudentially rational'). The fact that the belief is 'epistemically irrational' is a consideration the importance of which is seen as being overridden by prudential considerations (in something like the way that the supposed epistemic irrationality of a belief in God would be trumped by the prudential considerations of Pascal's Wager). This is not to say that self-deception *would* be the prudentially rational option in the circumstances, since we may indeed want to point out to such a person that in the long run he would be better off facing up to the truth.[13] It is only to say that making himself believe that *p*, on this intentionalist account, is something that *the subject* judges it best or reasonable to do despite the fact that this belief may be untrue or unwarranted. It is only to say that it seems reasonable from his point of view.

Again, the deflationist steadfastly denies point (iii), deeming it an over-intellectualization of the phenomenon. This denial is typically motivated by the thought that

[13] There may be a deeper way in which the self-deceiver might be accused of irrationality here if we take it to be rational and honorable to face up to the truth, and where we see the preference for comforting delusions as a sign of vice or weakness. In that further sense the intentional self-deceiver may be guilty of irrationality.

the idea of intentionally deceiving oneself is paradoxical, and also by the thought that the belief can be explained in simpler, less controversial ways.

(iv): 'Due to his efforts, Jones ends up in a condition where he believes that *p* and believes that not-*p* simultaneously'.

This next point concerns the *outcome* of the subject's execution of the self-deceptive strategy. It states that he ends up in a condition where he believes that *p* and that not-*p*, which may be termed the *contradictory belief condition*. Deflationism, on the other hand, is defined by its rejection of attributing any apparently paradoxical features to self-deception, and so deflationists deny that self-deception involves believing that *p* and that not-*p* at the same time. That they do so, indeed, follows from something we mentioned earlier, namely, that in the deflationists view, the self-deceiver doesn't draw the warranted conclusion that not-*p* to begin with. Since the warranted belief never gets formed at all, there is no question of the self-deceiver being in such a state for the deflationist.

The relationship between point (iv) and the preceding points, however, is as I've said previously not so straightforward, and some who endorse points (i) to (iii) deny that the contradictory belief condition necessarily follows from that and seem to reject this condition in their own accounts, while others endorse the contradictory belief condition, but reject the intentionality condition. Furthermore, the reasons why one might be drawn towards the contradictory belief condition are not singular. We will come across these in time. However, probably the most important reason for endorsing (iv) is that it is thought to account for a basic fact about self-deception, which is that self-deception essentially/typically involves tension or conflict in the mind and behaviour of the self-deceiver. Contradictory beliefs are then posited to explain the existence of this. As I said before though, we are going to leave discussion of (iv) aside for now until part II, and in part I we will take the debate between the

traditionalist and the deflationist to be one over whether intentionalism or non-intentionalism is right for answering the explanatory question.

In the next chapter we will look at empirical evidence showing the kinds of actions that are doing the explanatory work in Basic Scenario cases, whereupon we will look at intentionalist and non-intentionalist suggestions as to what the content of the intentions associated with these actions is.

# Part I: The Process of Self-Deception

## Chapter 2: Some Empirical Evidence on what Mediates Between the Desire and Belief, and its Interpretation

### *1. Introduction.*

To remind us of where we are, I have assumed that ordinary Basic Scenario cases are paradigmatic cases of self-deception. According to this view, self-deception involves being acquainted with evidence warranting the belief that not-*p*, but nevertheless believing that *p* because one desires that *p* (where one would have ended up believing that not-*p* had one not had that desire). And we are interested in establishing the nature of the explanatory connection between the desire and belief in these cases, which will show us the process of self-deception.

If we are concerned, then, with the question of *how*, in real-life Basic Scenario cases, one ends up believing that *p* because of a desire that *p* (i.e. the explanatory question), we may count ourselves fortunate that psychologists have set about demonstrating in experimental situations phenomena that conform, or that at least conform approximately (more on this later), to the Basic Scenario, and have also inquired into what they often call the *mediating processes*, that is to say, the processes mediating causally between the desire and the belief, which is of immediate relevance to answering the explanatory question. A good deal of this chapter, then, will be dedicated to outlining the fruits of this experimental research.

Is it being suggested, in that case, that the answer to the explanatory question is purely a matter for psychologists to ascertain? In particular, is it being suggested that all we need do is look to these experiments to see whether or not the mediating processes involve acts done

by the subject with the intention to deceive himself? What is the role here for the philosopher? I hold that the input of psychology for answering this question is significant. The experiments will show that the mediating processes in Basic Scenario situations involve actions of the subject (in particular, mental actions, such as reasoning), actions which are plausibly construed as intentional. However, psychologists don't themselves address the question of *what the intention is* with which these actions are done, and for that reason these results do not, as such, provide any ultimate vindication for either the intentionalist or the non-intentionalist positions. Thus once we have our hands on the results of these studies, I will consider it to still be an *open question* whether we should understand these actions in intentionalist or non-intentionalist terms.

In this chapter, the main aim is firstly, to identify the actions involved in these cases, and then secondly, to develop a plausible-sounding, *non*-intentionalist description of the intention associated with these actions, and compare it with the intentionalist one. However, though the non-intentionalist description of the intention is the one I wish to endorse, at this stage, I will only be putting these options on the table. I will not be giving reasons for thinking that one description should be endorsed over the other, a task that will be taken up in following chapters.

However, it is worth noting that if two people were to get into a dispute over whether these actions should be understood in intentionalist or non-intentionalist terms, it is unclear whether any further *empirical* investigation could resolve this issue for them. What seems to be required at this point is some investigation of a more *philosophical* character, which will include, in particular, answering some very general questions about the nature of intentional action. This investigation will be left for another chapter, and what I will ultimately argue is that on the assumption that the intentionalist's characterization of the intention holds, the unwarranted belief couldn't possibly have obtained. Therefore, by a process of elimination,

we should accept the non-intentionalist characterization of the intention presented in this chapter.

## *2. Experimental Demonstrations of the Basic Scenario.*

The kind of psychological experiments that we are going to look at are investigations into a phenomenon that is often put under the heading of 'motivated cognition', 'motivated belief', and 'motivated reasoning' (Mele, for instance, frequently uses this expression). Some clarification will be required, however, concerning how this expression should understood. The first point that may be made here is that there is a very general sense in which one could say that motivation (that is, desire) figures in the causal explanation of belief formation, a sense that would be too wide to capture the kinds of cases we are interested in here. Consider Johnny doing a maths exam. Johnny may have a strong motivation to get the answer to question 1; he desires to get a good mark. Without this desire, he wouldn't even bother to investigate the question, indeed, he mightn't even be in the examination hall. In analogous ways, we could say that a desire or interest of *some* kind is always necessary to motivate any kind of investigative activity, and in that sense, perhaps almost all beliefs are 'motivated', or caused or influenced by motivation. But when people speak of 'motivated belief' in the context that concerns us, they have in mind belief formed and maintained under the influence of a certain species of wants, namely, 'any wish, desire, or preference that concerns the outcome of a given reasoning task' (Kunda 1990: 480). This is what psychologists call *directional motivation*, and it denotes a desire, wish, want or preference concerning what is true of some issue, and it is often contrasted with *accuracy motivation*, in which a person has a desire to establish the truth about the issue, without caring about *which* answer is true. Johnny for instance, could not care less whether the answer turns out to be 8, -8, or 1,008. He

has no personal stake in any of those answers being right. He just wants to find the correct answer, whatever it may be. He has accuracy motivation, but no directional motivation, with respect to what the answer to the maths problem is. So what's often called a 'motivated belief' means a belief formed under the influence of a certain species of motivation: directional motivation.

It is in the examination of psychological studies into the influence of directional motivation on attitudes that we will find phenomena that conform, or conform approximately (as I will explain), to the Basic Scenario. Then we will be able to look at these studies for insight into the mediating processes. Consider the following well-known experiment by Ziva Kunda (1987, third experiment):

> A large group of undergraduate students read an article reviewing recent research which alleged that women who were moderate to heavy caffeine consumers were putting themselves at serious risk of developing fibrocystic disease, said to be associated in its advanced stages with breast cancer. Female high caffeine consumers–defined as those who drank three or more cups of coffee a day–were said to be at serious risk. Subjects later had to fill in a questionnaire which included questions on how convincing they found the article to be. It was found that the female subjects who were high caffeine consumers were more skeptical of the article than were male high and low caffeine consumers and female low consumers, all of whom found the article about equally convincing.

We can distinguish between two groups in this study. First, there are those who presumably felt something to be at stake for them personally in relation to the alleged link (the female high caffeine consumers). Call them the 'stakeholders'. And then there are those who we can

presume to have felt nothing to be at stake for them personally in relation to the allegation, or who were if not completely disinterested (they might have had loved ones who drink a lot of coffee) then at least relatively disinterested compared to the stakeholders (everyone other than the female high caffeine consumers). Call these the 'non-stakeholders'. Stakeholders, then, were found to be more skeptical of the article than non-stakeholders.

The question is: did the stakeholders here meet the conditions of the Basic Scenario? First let's consider whether the views of the stakeholders were *unwarranted by the evidence*. But how do we establish what belief the evidence warrants? One way of doing so is to take the judgments that competent and impartial judges of the evidence would make as an objective guide to what the evidence warrants. Then we could assume that the non-stakeholders are such judges, since they were impartial (having no stake in the issue) and were presumably reasonably competent to assess the evidence (being presumably intelligent third-level students). If this is a reasonable assumption to make, then if the views of the stakeholders deviated from that of the non-stakeholders, we could take the views of the stakeholders to be unwarranted by the evidence.

So the stakeholders apparently pass some of the conditions of the Basic Scenario, in that they were acquainted with evidence warranting one belief, but they formed a belief not warranted by that evidence, and deviating from the warranted belief in the direction of what they wanted to be true. But what was the *cause* of this deviation? If the culprit here isn't their desire that this belief be true, then the case will fail to conform entirely. So if, for example, the cause of the deviation could be attributed to purely cognitive factors, such as differences in relevant background beliefs or cognitive ability between groups, there would be such a failure. However, Kunda ensured that the difference in judgments between groups could not be attributed to differences in background beliefs (and presumably, cognitive ability).This was done by having the alleged ill effects of caffeine apply only to women, who presumably

had the same prior beliefs about caffeine as the male high caffeine consumers in the group (difference in sex does not seem like it would make one group have different prior beliefs about caffeine than the other). There was a deviation between the judgments of the stakeholders and these males that was the same as between the stakeholders and the other non-stakeholders. So we can presume that the stakeholders' deviation in judgment cannot be explained in terms of cognitive differences, and must instead be explained by the motivational differences between the groups. In other words, there is good reason to think that the stakeholders were more skeptical of the alleged link between caffeine and ill health *because* they desired for there to be no link, whereas were it not for that desire, they would have ended up making the same judgments as the non-stakeholders.

So the situation of the subjects in this experiment seems to conform to that of the Basic Scenario, insofar as they have an unwarranted belief, biased towards what they want to be true, which they have because of their desire for that proposition to be true. Or at least this conformity would be realized if it were true that the difference in attitude between the stakeholders and non-stakeholders was that the former *didn't believe there was a link* while the latter *did believe there was a link*. In actual fact, the difference was much more subtle than this. In the questionnaire designed to probe subject attitudes, Kunda did not ask subjects whether they believed or disbelieved the proposition that caffeine is linked to fibrocystic disease. She asked them, rather, to indicate on a 6-point scale how convinced they were of the purported link, where '1' meant 'not at all convinced' and '6' meant 'extremely convinced'. Stakeholders were on average a little less confident in the link than non-stakeholders (stakeholders' level of conviction averaged at 3 and non-stakeholders level of conviction averaged at 3.5). So strictly speaking, stakeholders didn't meet the conditions of the Basic Scenario (i.e. they didn't end up believing that *p*, because of a desire that *p*, while the evidence warrants the belief that not-*p*), though they did *approach* to satisfying these

conditions. In fact, the difference in attitude between stakeholders and non-stakeholders can't be represented in terms of outright belief in the proposition at all. Instead, we have to reach for some other descriptive tools, as with representing their attitudes in terms of their degree of conviction in the proposition. (It is worth noting that deviations between stakeholders and non-stakeholders of a limited magnitude such as this is the norm in studies into the influence of desire on belief).

However, it seems that this only shows that our phrasing of the Basic Scenario in terms of the belief that *p* and that not-*p* is too restrictive, since if we are prepared to accept that believing that *p* when the evidence warrants the belief that not-*p*, where this is caused in the appropriate way, is sufficient for being self-deceived, then we should be prepared to accept that having an *unwarranted degree of confidence* in a proposition, where that is caused in the same appropriate way, should also qualify one as self-deceived. For this would be just a less extreme variety of essentially the same phenomenon. It may seem like caviling to insist on this point right now, but it will be shown later that there are important theoretical advantages in rethinking self-deception in terms of unwarranted degrees of conviction in a proposition, rather than unwarranted beliefs in them (it will help us explain an important feature traditionally associated with self-deception that I will talk about later, namely, its 'tension').

Rethinking it in these terms also opens the way for us to consider some of the stakeholders in Kunda's experiment as being self-deceived. But there is no need to involve ourselves in these issues just yet, and for now I will mostly go along with the tradition in the literature of representing the attitude of the self-deceiver in terms of outright belief. For now we may just look on what was demonstrated in experiments such as these as being less extreme cases on a continuum with the kinds of cases represented in the literature, where the evidence warrants an outright belief in a proposition though the self-deceiver believes the

contrary. Consequently, we will be able to assume that investigation into the mediating processes between the desire and belief in these more subtle cases will be wholly relevant to answering the explanatory question, since it seems reasonable to suppose that the processes will be similar whatever the degree of deviation from what's warranted a subject's attitude may display. This is an assumption I will work with for now, at any rate.

### *3. Mediating Processes between Desire and Belief.*

In this section we will take a look at psychology's findings concerning the processes mediating between the desire and belief in Basic Scenario situations, and I shall offer a fairly unified account of what they are. One thing we will notice about them is that they involve *actions* (including mental actions), actions motivated by the desire that *p*. These actions display *bias*, and through that they result in the subject having an unwarranted belief in the likelihood of the proposition. The question of relevance, then, to the dispute between intentionalists and non-intentionalists will be whether these biased actions are being done intentionally (i.e. are intentionally biased), and with the intention of forming the unwarranted belief, and this issue will be discussed afterwards.

As we have seen in studies like that of Kunda's (1987), psychologists have found that when different people are presented with evidence alleging that some proposition is true, some of whom have a stake in that proposition (wanting it not to be true), and some of whom have no personal stake in the matter, these 'stakeholders' tend to believe in the likelihood of the proposition less than non-stakeholders. Initially the aim of these investigations was merely to determine whether it was true that such preferences as to the truth of the matter could bias attitude formation, and a consensus got reached that they can. However, other studies then went further by trying to investigate *how* it is that such desires influence their

beliefs. I will now present one such study which got results that are fairly representative of findings across the field. This will show us the kinds of actions that may be involved in producing the biased beliefs, and then we can address the question of the nature of the intention with which they are done.

Liberman and Chaiken (1992) replicated Kunda's 1987 study (with some variations) and included additional measures (in the questionnaire) designed to identify the processes underlying the formation of the biased belief. Subjects were told that they were participating in a study on 'nonscientists' understanding of scientific and technical information' (1992: 672). Then they were given an article to read purporting that research has shown that women who are moderate to heavy caffeine consumers are at much higher risk of developing fibrocystic disease, a serious ailment associated in its later stages with breast cancer. This time the subject pool consisted of women only, who were categorized on the basis of their coffee drinking behaviour, with 'low-relevance' subjects being those who drank no cups per day (making them non-stakeholders), and 'high relevance' subjects being those who drank 2 to 7 cups per day (making them stakeholders). After reading the material, they filled in a questionnaire about it.

In this questionnaire, subjects had to indicate their agreement with the statement that caffeine consumption causes fibrocystic disease and their belief in the importance for women to reduce their caffeine intake, both on a 9-point scale (their answers to both questions were averaged to form a single belief index). High-relevance subjects were found to be less convinced of the link than low-relevance subjects. Their scores averaged at 5.60, compared to 6.72 for low-relevance subjects (again, a subtle (yet statistically significant) deviation). Thus their conclusions displayed a bias towards what they wanted to be true, as in Kunda's study.

However this study went further than Kunda's by including additional measures designed to identify the processes responsible for these biased attitudes. For example, the

article contained four research reports on whether caffeine is linked to this disease. Of these reports, one was anti-link while the other three were pro-link. Methodological flaws or weaknesses were included in each report '[to] allow vigilant subjects the opportunity to be critical' (672). Subjects were then asked in the questionnaire to list any strengths and weaknesses they found in these reports, and their answers to this could be used to assess how the information was being treated.

It seems that if the biased attitudes are going to be caused by the subject's behaviour, then this behaviour will take one of two forms: a *fight* response, or a *flight* response. The former would involve trying to explain away or undermine somehow that unwelcome evidence, while the latter would involve trying to ignore it or turning attention away from it. The two processes that the experimenters tested for correspond to these two possibilities.

On one account of how such subjects end up with a biased belief, they do so by turning their attention away from any unwelcome evidence and towards evidence that would support their preferred conclusion (this is a flight response, and later we will see in chapter 4 that it is a common assumption among intentionalists that this kind of behaviour is operative in self-deception). The experimenters tested for such an explanation, which they termed a *defensive inattention hypothesis*. In the current context, if the subjects were to arrive at their biased belief by this route, presumably they would have avoided attending to, skimmed over, or tried forgetting about, the pro-link sections of the article and paid more attention to the anti-link section. This should then have predictable effects if afterwards they were asked questions about the details in the pro- and anti-link sections. If they avoided attending to and reflecting on the pro-link reports but not to the anti-link reports, we could expect that, if quizzed, they would get more questions right about the details of the anti-link reports than for the pro-link reports. The experimenters therefore included a recall test for the details of the reports to test for the defensive inattention hypothesis. Subjects were also asked for their

views on the strengths and weakness, and overall soundness, of each report. This measure was used to assess how much effort they put into scrutinizing the pro- and anti-link sections of the article.

From the evidence gathered, Liberman and Chaiken concluded that 'defensive conclusions were apparently not mediated by inattention to the message' (675), a finding which they say is consistent with those of other relevant studies. High-relevance subjects expended marginally more effort in reading the article overall than their low-relevance counterparts. Also, they apparently did not avoid reading the pro-link sections. The anti-link report was slightly better recalled, but this was attributed to the fact that it stood out as the only dissenting view among three pro-link views[14].

Liberman and Chaiken found evidence for a different mediating process to defensive inattention, which they dubbed *biased systematic processing*. High-relevance subjects were discovered to have regarded the anti-link report as superior to the pro-link reports. Importantly, whereas low-relevance subjects found an equal amount of weaknesses in pro- and anti-link reports, high-relevance subjects listed 'significantly more' weaknesses in the pro-link reports than in the anti-link report, and listed *less* weaknesses in the anti-link report compared to low-relevance subjects. As the authors conclude, '[c]ompared with low-relevance subjects, high-relevance subjects were less critical of those parts of the message that were reassuring and more critical of those parts that were threatening' (675).

A study by Lundgren & Prislin (1998) found deviations in judgments between stakeholders and non-stakeholders caused by similar processes. Groups of students had to give their opinion on whether there should be a 30% tuition fee increase in their university.

[14] Another group of subjects were given a different 'low threat' article in which there were three anti-link reports and one pro-link report. Here, details of the one pro-link report were recalled slightly better. Thus there is a general tendency for a section of an article to be recalled better simply because it stands out from the other content of the article by representing a dissenting view.

For one group, they were told the proposal was for 10 years time (making them non-stakeholders), and they were motivated to form an accurate judgment (by being told that their reasoning abilities and competence would be examined). For another group, they were led to believe the tuition fee proposal was for next Fall, and they were told that their views would help the board of regents make their decision (making them stakeholders). Both groups were given access to information and arguments on the increase. Non-stakeholders were found to arrive at a neutral position on the increase, while stakeholders ended up strongly opposing it. Measures added in the experiment gave insight into the differences in reasoning between both parties.

From making pro- and anti-tuition fee increase information available to these subjects and seeing which, in a limited time, they go for, and also by having both groups do thought-listing exercises (writing down their thoughts) as they reason and analyzing these thoughts, the authors could investigate the way that they assessed the issue. Both stakeholders and non-stakeholders were found to have put equal effort into their assessments. However, they found that the stakeholders 'focused almost exclusively on information that supported their preferred position in both searching for information and thinking about the tuition increase [i.e. in generating their own arguments]', and they displayed an 'avoidance of antagonistic information'[15] (1998: 721). On the other hand, non-stakeholders 'thoroughly examined all of the available information, engaged in extensive and equally balanced thinking whereby they elaborated external information, and also generated novel arguments' (720).

These studies and others like them suggest that subjects end up with their biased beliefs by trying to construct what to their eyes seems like a justification for their desired

[15] It may sound like the assertion that subjects avoid antagonistic information contradicts the idea that their main strategy is to refute that information, since the attempt to refute the information would involve considering it and so not avoiding it. But there is no contradiction here. What is meant is that subjects avoid or neglect seeking out *new* antagonistic information. With antagonistic information that they have *already* encountered or are *forced* to encounter, subjects rely on refutatory strategies.

position (Sanitioso *et al* 1990: 229. Kunda 1990: 483). In doing this, they deal with the unwelcome evidence by basically adopting a fight response: taking a *hypercritical stance* towards it, making a concerted effort to discredit it, while simultaneously seeking and being uncritical towards any welcome evidence (including the considerations they adduce against the unwelcome evidence). In other words, *they one-sidedly put their energy into seeking or thinking up favourable considerations, while neglecting to seek or think up, to a comparable extent at least, unfavourable ones*. Thus they end up being exorbitantly critical of evidence against *p*, and overly credulous or not critical enough towards evidence for *p*. These results are consistent with numerous other studies showing that people tend to selectively seek information supportive of their desired view when that view is threatened (e.g. Ditto *et al* 2003; Ditto & Lopez 1992; Frey & Stahlberg 1986; Holton & Pyszczynski 1989; Pyszczynski *et al* 1985; Wyer and Frey 1983.). For instance, Frey found that subjects led to believe that they had performed poorly on an intelligence test showed a preference for reading articles critical of intelligence tests over articles supportive of them when given the choice to read either (Frey 1981). Interestingly, though perhaps unsurprisingly, in the Lundgren & Prislin study, when subjects were asked to indicate whether they believed their view was a logical one, stakeholders indicated that it was, no less than non-stakeholders, thus showing their lack of appreciation of the biased nature of their own view[16].

That this behaviour counts as *biased* should be fairly clear. On one standard meaning of 'biased', it implies not giving each side of an argument its due, and being unbalanced in one's evaluation of an issue, usually because of a special interest that one has. The relevant sense of bias I have in mind is here articulated by Annette Barnes:

[16] I say 'unsurprisingly' because it is difficult to conceive of how one could hold a certain view while also believing that it is *not* a reasonable view. Moreover, it shouldn't be surprising that people can gather evidence one-sidedly while not appreciate that they are being biased. This kind of thing also happens in 'cold' cases (i.e. cases where a subject evaluates whether *p*, with no desire concerning whether *p*) of the confirmation bias. See (Wason 1960).

> While some beliefs are not arrived at after deliberation and decision, other beliefs are. With regard to some of these latter beliefs, a person is epistemically responsible in acquiring them only if, in his deliberations about what to believe, he imposes on himself something akin to the adversary principle of considering both sides. In the public arena, rational thought has as its ideal an impartial judgment arrived at after hearing non-corrupt representations of opposing sides. Different persons represent the opposing sides, and judgment is rendered by a person independent of either side. In the private sphere, a person, as a rational believer, is sometimes expected to function analogously: to be not only an impartial judge who hears both sides but a non-corrupt representative of first one and then the other side (Barnes 1997: 85-86).

This is exactly what these stakeholders tend not to do. Because the stakeholders are more interested in finding fault with the unwelcome evidence, and finding support for the welcome position, than in giving each side its fair hearing, their efforts are biased in this sense, resulting in the generation of a one-sided body of considerations from which they make their judgment, compared to the more balanced body of considerations that would be generated by the thinking and evidence gathering of impartial judges. Thus we can easily understand how this behaviour would render stakeholders' conclusions less reasonable. But why should stakeholders take seriously their biased body of considerations as an adequate basis for a reasonable judgment? Lundgren & Prislin suggest it is 'because they [mistake] extensive processing for objective processing' (1998: 721). In other words, because they put much effort into examining the issue, they assume they have examined it thoroughly and well.[17] But

[17] It should not come as a major surprise that people can one-sidedly try to refute or confirm a proposition and assume that they have examined the issue objectively, since this same pattern is evident even in cold cases of hypothesis testing with the confirmation bias, where there is no preference regarding what's true.

this is not so, since their examination has been biased towards collecting considerations that would support their preferred position, without equal effort being put into collecting contrary considerations. As L.S. Newman notes, 'extended and careful thought, generally considered to be a virtue, can actually be associated with more and not less bias' (Newman 1999: 60. Also see Kunda 1990: 491).

### *4. Philosophical Significance of the Empirical Findings.*

We have by now looked at some examples of studies probing the mediating processes between the desire and belief in Basic Scenario cases. What is of particular philosophical significance in the results of these studies is that the processes involve kinds of *actions* carried out *by the subject* (reasoning, evaluating, criticizing, searching for evidence, etc.), actions which display bias. This makes these cases conform to what we would expect of a case of self-deception as something perpetrated by *us*, and not by a mere mechanism of some sort, triggered by the desire. As Scott-Kakures puts this point, '[t]o be self-deceived, the self-deceiver must himself take an active role in his deception' (2002: 591, also see Michel and Newen 2010: 742). So these studies seem to vindicate our assumption that the investigation of ordinary Basic Scenario cases counts as an investigation into self-deception.

Furthermore, if we return now to the My Lai Massacre story mentioned in the introductory chapter, we can see that this was evidently the case in this situation. Subjects who encountered unwelcome evidence that American troops committed an atrocity–evidence of a type that would be 'ordinarily granted credibility'–were seen to have engaged in *rationalization*, that is, in efforts to *explain away* this evidence ('The story was planted by Viet Cong sympathizers', etc.), actions which may plausibly be construed as intentional. This may very well have played a major role in supporting our inclination to categorize them as

cases of self-deception at the time. We might have extracted an 'action condition' from these cases and supposed it to be necessary for self-deception, and put it down as a constraint that any answer to the explanatory question for Basic Scenario cases must meet if they are to count as self-deception, according to which the desire in self-deception must cause the belief by motivating actions which result in the belief. However, at the time I didn't want to do this because it would have immediately excluded a certain approach that one might take to the explanatory question, one which at that early stage ought not to have been excluded. This was the approach of non-agentive deflationism.

This approach should not have been excluded back then because it is apparently a live position in the literature. Ariela Lazar is one notable proponent of this approach (1999). Lazar tries to give a more prominent role to the emotions, rather than desire, in her positive account of self-deception. She mentions cases where emotions and moods affect cognition without the mediation of action, as when one perceives the world and thinks of the future more positively or negatively depending on whether one is upbeat or depressed, or when one's anger with someone causes her to view him distortedly as a self-centered and inconsiderate person, rendering salient all his faults. According to Lazar, these emotions exert an immediate influence on cognition here since it is apparent that they don't *cause us to act* in some way, with the result that we end up with an overly positive or negative view of the world. However, it is unclear what exactly she is doing in mentioning these cases, since she doesn't explicitly promote them as being actual cases of self-deception (and we have little pre-theoretical inclination to regard them as such), and may just be using them to illustrate how emotions affect cognition in an immediate way, with the suggestion being that similar processes may be operative in self-deception. Nevertheless she says '[t]he assignment of a central role to the emotions in the formation of self-deceptive beliefs is largely incompatible with the view of self-deception as an action' and that these emotions can affect our beliefs

'immediately and in a way which, to a high degree, is not subject to our control' (1999: 282). Mark Johnston (1988) seems to advocate a similar position, holding that the desire in self-deception operates by activating a 'mental tropism', or a 'subintentional mental process' which produces or sustains the unwarranted belief. The implication here seems to be that the desire causes the belief not by motivating intentional action, but by means of some non-rational mechanism or process not subject to our voluntary control which causes/sustains it, and on this view there would also be no self-deceptive actions performed by us.

These philosophers seem to have been drawn to this position in the belief that ordinary Basic Scenario cases, which they assume are self-deception, are best explained purely in terms of the non-rational effects desires may exercise on cognition. However, the evidence we examined on Basic Scenario cases does not seem to vindicate this belief, since it suggests that subjects end up with their biased belief because they are *acting* in a biased way. And even if the evidence suggested a purely non-agentivist answer to the explanatory question, for certain cases at least, this might only cause us to review any inclination we might have had to classify those cases as self-deception. For we may wish to endorse the intuition that the process of self-deception must involve the subject's actions motivated by the relevant desires and emotions, if we are to have a genuine case of self-deception, and if that were so, non-agentivist accounts simply won't be contenders in this debate at all. In that case, cases where the mediating process between desire and belief is purely non-rational will simply be a case involving a 'deviant causal chain'. Either way, it seems that we may leave non-agentive non-intentionalism behind us, and take the real debate to be between agentive non-intentionalism, and intentionalism.

What we know so far, then, is that there are human actions mediating between the desire and the belief in Basic Scenario-type cases. These actions, broadly speaking, are what we might call *evaluative behaviours* (scrutinizing considerations, searching for new

considerations, etc.), in that they are behaviours which one undertakes typically for the sake of evaluating evidence, or a proposition, for validity or truth, and furthermore, they are behaviours which happen to be biased. Now it is difficult to imagine how these actions could be done by the subject, but not be done with any intention or for any kind of reason. For evaluative actions are goal-directed: directed at the goal of determining truth, and are typically engaged in for that reason. These kinds of actions don't seem like the kind we could assimilate to any recognized category of non-intentional voluntary actions, such as idly performed actions (e.g. twiddling) or expressive actions (e.g. smiling)). So in order to explain their existence, we must articulate the intention with which they are being done.

However, for all that has been said so far, this empirical evidence may seem to be neutral between intentionalism and non-intentionalism. For the crucial question relevant for adjudicating between these theoretical approaches will be: *what are the intentions with which these actions are done?* Are they done with an intention to deceive oneself, or with another intention? And this is a question that is not addressed by the psychologists who conducted these studies. Let us now take a look at the different kinds of characterization of the intention that would be offered by both intentionalism and non-intentionalism.

## *5. Intentionalist and Non-Intentionalist Descriptions of the Intention.*

We have seen that the culpable actions involved in paradigmatic cases of self-deception consist of seeking considerations which may undermine the unwelcome evidence, and perhaps also considerations that are independently supportive of one's preferred conclusion that *p* (all of which we may simply call 'favourable considerations'), while neglecting to

seek, or to seek with an equivalent effort, unfavourable considerations. An intentionalist take on this would be as follows. For the intentionalist, the subject intentionally executes a deceptive strategy with the intention of deceiving himself. And the empirical data here seems to show what the strategy is: the subject is intentionally seeking favourable evidence/considerations with the intention of *making himself have the belief* that *p*, a belief he initially knows to be false/unwarranted. The suggestion here would be that he is intentionally seeking evidence, and evaluating the evidence already in his possession, in a one-sided and biased way, in the hope that this will lead him to form this belief. If successful, he will be taken in by this one-sided evaluation process, and will form the belief that *p* as planned.

This interpretation may seem prima facie paradoxical. How, one may wonder, could one *intentionally* act in a biased way and be hoodwinked by this activity? However, intentionalists have a number of ways of trying to explain away the initial sense of paradox, including making appeals to unconscious intentional action, and theories of 'mental partitions', all of which will be examined in due course. There are various reasons why philosophers would adopt the intentionalist interpretation of these cases. One of these will be the sole topic of chapter 4. However, the most common reason seems to me to be one which goes unsaid. Here it is important to note that, traditionally at least, people who think self-deceivers must intentionally deceive themselves have not tended to arrive at this conclusion from examining real life cases. Most have taken this to be a conceptual truth, which can be seen simply from *analyzing the notion* of self-deception. However, many of these people are also convinced that self-deception is a real phenomenon. So when a case is presented which intuitively strikes us as one of self-deception, their instinct has been to view it in line with their analysis. These a priori arguments for intentionalism will be examined in chapter 3. For now, however, let's just take note of this as the intentionalist interpretation of these cases, and move on to the proposed non-intentionalist interpretation.

For the non-intentionalist, then, these actions are not intentional under the same description 'seeking considerations/evaluating in a biased way', since on this view, we don't see the subject as intentionally acting in an epistemically untoward, deceptive way. Alternatively, all we need to say to account for the subject's biased behaviour, call him 'Jones', is the following: After the unwelcome evidence that not-*p* is encountered, Jones finds that a previous assumption of his is threatened with being false. Because of his strong desire that *p*, this prospect is a source of great distress for him. Because of this, Jones has a *heightened interest* in any evidence/reasons there might be that would invalidate that evidence and remove the threat, for the reason that *it would put his mind at ease to have p confirmed*. Jones hence becomes anxious to find reasons that may invalidate that evidence, which, again, if found would allow him to *rest assured* that *p*. Anxiously desiring to find reasons to discount that evidence, he subsequently acts for the sake of satisfying this desire. His intention, then, is simply to fulfill that desire. That is to say:

> *Jones acts with the intention of finding any p-confirming/supporting evidence/considerations*.

…where considerations which undermine the unwelcome evidence for not-*p* are supposed to count as '*p*-supporting', in at least the weak sense of removing a threat to the assumption that *p* (we may introduce the term '*p*-favourable' to cover the notions of '*p*-confirming' and '*p*-supporting').[18] Now it may be observed that for the intentionalist, the subject acts with this intention also. On that view, Jones intends to find *p*-favourable evidence too, but the difference is that here he is doing this in the attempt to satisfy his further intention of making himself believe that *p*. What I am saying is that the intention to find *p*-favourable evidence

[18] Such considerations may not necessarily *positively* support *p*. But self-deceivers will also often be spurred into seeking independent considerations to positively support or confirm that *p*.

needn't be seen as a sub-intention in this larger intentional project for it to be intelligible to us why Jones is acting in this way. For simply in light of the fact that he desires desperately that *p* we can understand why Jones would be driven to seek *p*-confirming/supporting evidence, simply because his finding it would assure him that what he wants to be true is true, thus putting his mind at ease. This also allows us to see the desire that *p* as playing not merely a causal, but a motivating, rationalizing role in the explanation. In light of it, we understand why having this intention fulfilled would be so attractive from that agent's perspective.[19]

The evidence search could be either physical or mental. To illustrate the former, Mele describes a case of a historian of philosophy who hopes that her favourite philosopher, Plato, holds the same view as she holds on some matter. 'Consequently, she scours the texts for evidence of this while consulting commentaries that she thinks will provide support for the favoured interpretation. Our historian may easily miss rather obvious evidence to the contrary, even though she succeeds in finding obscure evidence for her favoured interpretation' (1997a: 94). Here, the search involves physically looking through various texts, which she believes might turn up evidence that Plato held this view. A lot of the time though, the action will be purely mental, being a matter of trying to think up of reasons in support of the preferred conclusion (in cases involving rationalization). In general, the self-deceptive action will be some kind of searching, done with the intention of finding.

Hopefully this provides a coherent non-intentionalist picture of what the self-deceiver is up to here which we can empathize with and find plausible. He encounters evidence seeming to undermine a cherished assumption, and becomes anxious to find any weaknesses in that evidence, so that he can rest assured that *p* is true. Again, Jones is motivated to do this

[19] Note that it is not the case that on this view there is no role for the emotions in self-deception, but only for desires. Desires and emotions are intimately connected, and indeed, I would say that having a desire entails that you are disposed to feel certain emotions in certain circumstances. For instance, desiring that *p* entails that if you discover that not-*p*, you will have a tendency to feel negative emotion (a tendency which *may* be blocked by something). Clearly such negative emotions are playing a central role in the process of self-deception here.

because it would very much put his mind at ease to have the unwelcome evidence invalidated or to have *p* confirmed, given that he desires so much that *p*. This is why he searches vigorously for, and seizes upon, *p*-favourable evidence. Moreover, it should be clear that this intention cannot be described as an intention *to deceive himself*. For Jones to be intending to deceive himself here, he would have to be acting with the intention of *making himself believe* that *p*. That is, he would have to be concerned with installing an attitude in himself, regardless of its falsity. But on this view, Jones' concern is not with his own doxastic condition, but is 'outwardly directed'. He is, once again, anxious to put his mind at ease that it's *actually true* that *p*, and so *acts in the hope, and with the intention, of finding any considerations that would confirm this, not in the hope of making himself form a belief*. And this intention is sufficient to explain the subject's behaviour in these cases.

Note also the use of the word 'any' in the statement of the intention. This, I believe, is preferable to simply saying 'acts with the intention of finding *p*-confirming/supporting considerations', which may be considered paradoxical for the following reason. It is often held that a precondition for having an intention to *v* is that you believe you can *v*. Thus, if Jones intends to find evidence confirming that *p*, then he must believe that he can find such evidence. But then he must believe that such evidence that would confirm that *p* exists for him to find, and wouldn't that be tantamount to believing that *p* is true?

Of course, we don't want to imply he's already convinced that such evidence exists, for then why all the anxiety? Clearly, the anxiety of the self-deceiver must be explained by supposing uncertainty at the outset over whether such evidence exists. But is saying that he intends to find such evidence compatible with this? Consider this parallel case. I go into a shop because I want a particular magazine. I'm not sure that they will stock it, but I'm going to find out. Can we say that I'm going in there with the intention of finding the magazine? If this sounds wrong, then at least we can say that I intend *to find it if it's there*, or to *try to find*

*it*, or to *look for* it, or that I am acting *in the hope* of finding it. I have tried to formulate the self-deceiver's intention in a similar manner, hoping that the word 'any' conveys an initial lack of certainty that he will find what he wants to find. He acts with the intention of finding any *p*-favourable evidence there might be.

## *6. Further Elaboration of Differences Between the Intentionalist and Non-Intentionalist Accounts.*

In this section I would like to elaborate on some points relating to the non-intentionalist view of these cases, under a number of headings, and trace some of the important respects in which it differs from the intentionalist one.

*Practical reasoning*. We said that for the intentionalist, we can construct a practical syllogism for the self-deceptive actions (along the lines of: 'it would be desirable to believe that *p*, by doing *V*, I would best be able to believe that *p*, so I should do *V*'). The non-intentionalist will not attribute any such practical reasoning to the self-deceiver. But if so then what kind of practical reasoning may the non-intentionalist associate with self-deception? In general, we could say that the self-deceiver's culpable actions are seeking-type actions and that seeking in general is done with the intention, or with the hope, of finding. The practical syllogism associated with the intentional action could then be spelt out as follows.

- It would be desirable to find evidence confirming that *p*.
- By doing *V*, I could perhaps find evidence confirming that *p*.
- I should do *V*.

What *V* is will then differ from case to case. In the example taken from Mele mentioned above, *V* was the act of scouring various texts, which the subject hoped might throw up evidence for her favoured interpretation of Plato. A lot of the time though, *V* will be mental actions of trying to think up of reasons and considerations to discount or explain away the unwelcome evidence.

We should note that because, as I have said, the intention given in the non-intentionalist account is included in the intentionalist account, where it is a sub-intention in the project of intentional self-deception, the practical reasoning included in the non-intentionalist account will also feature in the intentionalist account. Presumably, it would be mentioned in the elucidation of the self-deceptive strategy, which features in the intentionalist's practical syllogism given above.

*On whether the warranted belief is formed or not*. Another important way in which this non-intentionalist account differs from the intentionalist one is that according to this view, Jones did not, on first encountering the unwelcome evidence, draw the unwelcome, warranted conclusion to begin with. Intentionalists portray the self-deceiver as correctly evaluating the evidence and realizing its true import to begin with, and *then*, subsequent to that, as proceeding to deceive herself. But for the non-intentionalist it is *during* this evidence-evaluation stage that the self-deceptive behaviour happens. The non-intentionalist need only say that when Jones encountered the evidence, he recognized that it *suggested* that not-*p*, and his subsequent biased scrutinizing was part of his effort to evaluate the validity of the suggestion. People don't necessarily draw conclusions immediately after encountering evidence, but often need to go through a process of evaluating it, and for the non-intentionalist, it is during this process, not after it, that the self-deceptive behaviour occurs.

*Awareness of bias*. Furthermore, the non-intentionalist here will claim that although Jones is evaluating the issue in a biased way, he does *not appreciate* the fact that he is. This

arguably follows from saying that he is not doing it intentionally under that description. It also further differentiates the non-intentionalist account from the intentionalist one, since one might think that on the latter account, because Jones is acting intentionally under the description 'acting in a biased way', he must be aware that he is acting in a biased way. Indeed, this idea is the source of the claim that the idea of intentionally deceiving yourself is paradoxical, and it will be discussed in detail in chapter 5.

This may raise a concern about this non-intentionalist account, however, since one might think it implausible that one could be acting in a biased way like this, without appreciating that one is. However, I believe that it should be uncontroversial that this can occur. If we find this difficult to understand we should bear in mind that a similar phenomenon occurs in 'cold' cases of the confirmation bias, when people assess whether *p*, without desiring that *p* or not-*p*, by one-sidedly seeking confirmatory evidence for *p* and without seeking refutatory evidence against *p* (e.g. see Wason 1960). Apparently, people have a common tendency to evaluate in this way, independently of motivational factors, without it dawning on them that they are doing anything problematic. If this kind of thing happens in 'cold' cases, it should be no surprise that it can happen in 'hot' cases too. Certainly then, there will be no need to suppose that self-deceivers have to *make* themselves unaware of the fact that they are being biased; such unawareness is simply the default position.

The appreciation of whether one is evaluating or searching for evidence in a balanced or a biased way should not be regarded as something that comes with such activities automatically. Rather, this is something that may have to be won with special effort, with the effort of exercising of a kind of second-order thought or awareness (i.e. where we reflect on our thinking process, noting if it is proceeding, or has proceeded, in a balanced manner), and this is something that we may be inclined to neglect to do (especially when emotions run

high). And as Lundgren & Prislin suggested (1998), subjects who energetically try to refute unwelcome evidence may mistake the effortfulness they put into their evaluation with overall thoroughness.

*Irrationality.* Let us now discuss the sense in which the self-deceiver is irrational on this view. Earlier when discussing the intentionalist view, we saw that some intentionalists distinguish between prudential rationality and epistemic rationality, and while they see the self-deceiver as being rational in the prudential sense (doing what they believe is necessary to achieve their goal of reducing their anxiety), they are nevertheless guilty of 'epistemic irrationality' because they end up holding a belief which is unwarranted by their evidence. Let's take a look at the sense in which the self-deceiver counts as irrational on the non-intentionalist view.

For the non-intentionalist, the self-deceiver is also irrational in the sense of holding a belief which is unwarranted by her evidence. However, we may also say that the self-deceiver *behaves* in an irrational way, in a way which violates certain important epistemic norms. We expect of an ideal or rational judge of the evidence that she looks at the pros and cons, strengths and weaknesses, of each side of the argument. She would be one who acts in accordance with what Barnes called the 'adversarial principle'. She would therefore be expected to have the intention to collect a balanced body of considerations before making her mind up. The self-deceiver, however, looks for evidence for *p* in the hope of having *p* confirmed, and the idea of collecting a fair and balanced body of considerations is not on his mind. In that respect, he deviates from what we would expect of a rational judge, and he is almost guaranteed to end up with a one-sided and hence *biased* body of considerations. He will end up overlooking considerations that he shouldn't overlook, and will fail to subject things to proper scrutiny. This contributes to the self-deceiver being irrational.

It may be important to note, however, that seeking evidence in a biased way is not necessarily irrational. Consider someone in a formal debate, or a lawyer at trial, for instance, tasked with defending the position that *p*. They may put much more energy into finding reasons for *p* than for not-*p*. But clearly they are not being irrational because of that. Why is this? Partly, I would say, it is because they are not under any illusions about the fact that they are being biased. For that reason we expect they will not be as credulous towards their position as their oratory might suggest. We think that the debater and lawyer are being to some degree insincere, an insincerity that is not just allowed for but expected in the circumstances, all of which is common knowledge between them and the other involved parties[20]. The self-deceiver, on the other hand, *is* under an illusion about the nature of his evaluation. In the heat of the moment, he doesn't have the presence of mind to realize that he has been biased or unfair, and rushes to a conclusion after having come up with some favourable evidence. This lack of knowledge is essential for a charge of irrationality to hold, as it seems necessary if the subject is to be taken in by his own biased behaviour.

Another important reason why the debater and lawyer aren't irrational has to do with their goals. Here we should note that it is with *respect to the goal of establishing truth* that the actions of searching for evidence/reasons in a one-sided way are judged inadequate. However, the lawyer and debater, who both act in a biased way, have goals other than that of establishing truth (their goals are those of defending a client, and of defending a position respectively), and their biased actions don't appear to be so inadequate relative to those goals. Thus the fact that these subjects are engaged in a kind of behaviour which is less than optimal for the purposes of establishing truth can't be a reason to criticize that behaviour, since establishing truth is simply not their goal.

[20] Though debaters and lawyers can end up convincing themselves of the position they are assigned to defend if they lose sight of the fact that they have been forced to be one-sided, and at that point they may be charged with irrationality.

This puts into contrast the self-deceiver as seen on the intentionalist and non-intentionalist accounts. For the intentionalist, as with the lawyer, the subject is behaving in a way that is inadequate for the task of establishing truth. But since the subject's goal in doing it is not to establish truth (rather, his goal is that of producing a belief), he can't be charged with irrationality in that respect. Nevertheless, the intentionalist still sees this as a phenomenon of irrationality, since the result of the actions is that the subject believes that *p* when her evidence warrants the belief that not-*p*.

However, for the self-deceiver on my non-intentionalist account, the subject *is* primarily concerned with the question of whether *p*, and so she *is* evaluating and searching for evidence with the goal of establishing truth. But relative to that goal, this behaviour is not up to the task, a point which she is hardly incapable of appreciating. So the activity of the self-deceiver on this account is not the exhibition of prudential rationality that it is for the intentionalist.

## *7. The Relation of this Proposal to Others in the Philosophical Literature: Mele's Processes of Self-Deception.*

I intend for this to be a fairly comprehensive account of the behaviour supporting self-deceptive beliefs. However, it will be important to consider how this proposal relates to others made in the literature, and here I would like to consider in particular the self-deceptive processes described by Alfred Mele. Mele mentions a number of processes that operate to produce motivationally biased beliefs that may seem disparate and disunified. But there may prove to be more unity to them than first meets the eye, and I think that many of them can be seen as instances of the kinds of actions seen above. This will also show us what I mentioned earlier: that although when Mele gives a general characterization of his approach, he

sometimes portrays himself as a non-agentive deflationist, when we look at the details of his view, we can often see him giving examples of cases where self-deceivers end up self-deceived because of their own actions, actions motivated by the desire that *p*. However, Mele seems to discuss other, possibly non-rational ways in which desires can influence and bias cognition. I wish to discuss these and emphasize that they may compliment the motivational explanation I wish to develop here.

One process mentioned by Mele, *selective evidence gathering*, is when 'Our desiring that *p* may lead us both to overlook easily obtainable evidence for not-*p* and to try to find evidence for *p* that is much less accessible' (2001: 27). This is quite obviously an instance of the kinds of behaviour seen above.

Another process he mentions is *negative misinterpretation*, which is when '[o]ur desiring that *p* may lead us to misinterpret as not counting (or not counting strongly) against *p* data that we would easily recognize to count (or count strongly) against *p* in the desire's absence' (2001: 26). The example is given of Don who has his paper rejected from a journal, and who forms the belief that it was wrongly rejected because the reviewers 'misunderstood a certain crucial but complex point'. Also, *positive misinterpretation* is when 'Our desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence' (2001: 26). Mele illustrates this with the case of Sid who interprets the rebuffs of a woman he's infatuated with as a sign of her playing hard to get. We can also understand the reasoning of Don and Sid, however, to be instances of one-sidedly seeking reasons to support one's preferred view. Don finds a reason to think that the reviewer got it wrong, a reason he doesn't scrutinize, while Sid finds a reason to think that some evidence is favourable to his desired conclusion rather than unfavourable, and accepts it unthinkingly. They both search for and seize upon welcome hypotheses that disarm the unwelcome import of the evidence, and they fail to consider alternative, less welcome

hypotheses, or to subject the welcome hypotheses they entertain to appropriate critical scrutiny.[21]

Another process Mele mentions is *vividness of information*. He says that 'motivation can increase the vividness or salience of certain data. Data that count in favour of the truth of a hypothesis that one would like to be true might be rendered more vivid or salient given one's recognition that they so count' (2001: 29). This salience effect, however, may be a consequence of the fact that one is concertedly looking for welcome evidence, since the very fact that one is looking for something tends to make it more salient than other things (think of looking for something in a cluttered drawer; one's attention skims over, and one barely notices, unwanted items, and lands straight onto the wanted item).

The biased reasoning and evidence gathering of the self-deceiver would result in him being in possession of a biased body of considerations from which to draw a conclusion, and this would probably be enough in itself to lead to a biased belief. However, this may not explain other things about stakeholders, such as the fact that they apparently are inclined to *jump to a conclusion* on the basis of those considerations in too quick a manner, or assign *improper evidential weight* to those welcome considerations. To explain this we may need recourse to a theory which has been developed by a number of psychologists and discussed by Mele, who calls it 'the FTL model' (an acronym referring to the work of psychologists J. Friedrich, Y. Trope and A. Liberman).

The FTL theory uses the notion of a 'confidence threshold', which refers to the quantity and quality of evidence required to convince one that something is true. This theory proposes that confidence thresholds may vary from case to case depending on what the expected costs are of *falsely* believing the relevant proposition. The less costs expected to be associated with falsely believing that *p*, the lower one's confidence threshold regarding it (i.e.

[21] The other process he mentions along with these, *selective focusing/attending*, is the defensive inattention hypothesis that has not been found to be operative in these cases.

it will take evidence of a relatively lesser quality or quantity to make one believe it), and the more costs expected to be associated with falsely believing that *p*, the higher one's confidence threshold (i.e. it will take evidence of a relatively higher quality or quantity to convince one of it).

Regarding some hypothesis *p* which a subject attempts to evaluate, the costs associated with wrongly settling the question in favour of *p* and the costs associated with wrongly settling the question in favour of not-*p* may be 'asymmetric' in magnitude. The subject will then be most concerned with avoiding the costlier error, and this will mean that he will have different thresholds for coming to believe that *p* than for coming to believe that not-*p*. If believing that *p* falsely would be very costly, then his threshold for believing that *p* will be high, and it will take relatively good evidence before he settles on that conclusion. Consequently, this subject may be biased towards concluding that not-*p*. He will be disposed to jump to the conclusion that not-*p* more easily than she would to the conclusion that *p*.

For instance, compare how a person, *S*, might evaluate whether she has some desirable character trait (likeability, moral virtue, intelligence, competence or talent etc), compared to how an impartial judge might do so. If *S* were to believe that she didn't have this trait when she did, then she would pay for this psychologically with a decrease in her self-esteem and confidence, and might discount her ability to do things that she is in fact capable of doing, hence missing opportunities. Whereas if she were to believe she did have this trait when she didn't, there would perhaps be no comparably serious costs to be incurred (indeed, perhaps there would be overall benefits). This asymmetry may mean that *S* has a lower confidence threshold for believing that she has the desirable trait, i.e., it will be easier for her to become convinced of that. On the other hand, her confidence threshold for believing that she does not have the desirable trait will be relatively high; it will take more evidence, or evidence of a better quality, before she becomes convinced that she lacks the desirable trait.

She will then be disposed to be biased towards believing that she has this trait. For another person who knows her and who is an impartial observer with respect to her on the other hand, there may be no particular costs associated with falsely believing that she has the trait, or doesn't have the trait, and so there may be no imbalance in his confidence thresholds for reaching either conclusion. These 'mechanisms', then, can explain why self-deceivers have a tendency to jump to the welcome conclusion on the basis of some welcome considerations which would not be sufficient to convince others. They can explain why they sometimes assign improper evidential weight to welcome considerations.

This model also predicts and explains cases of so-called 'twisted self-deception' where one unwarrantedly convinces oneself that some unwelcome proposition is true. Why, for instance, does the hypochondriac become convinced that he is seriously ill on the flimsiest of grounds? Using the FTL model, we could ask what costs the subject associates with wrongly believing either that he is or is not seriously ill. If he wrongly believed that he was seriously ill, this would surely bring costs in the form of unnecessary distress, which may tend to raise his confidence threshold for believing that. However, he may associate much greater costs with wrongly believing that he is not seriously ill. For if he took himself to be well when he wasn't, then he would miss the opportunity for remedial action when remedial action is urgently needed.[22] According to the FTL model, this will mean that it may take a lot to convince him that he is well, and not much to convince him that he is not well. This is why he jumps to the unwelcome conclusion so hastily.

What kind of an explanation is the FTL model suggesting? Is it a form of rational explanation? Would this be to suggest that the subject actually considers to himself what the

[22] As Scott-Kakures notes, what this model suggests is that what he calls 'unwelcome believing' cases 'should be pronounced only where the feared eventuality is regarded (rightly or wrongly) as in some way remediable, where there is some perceived degree of control with respect to the threat (2000: 365). Were remedial action not possible in such a case, for instance, then there would be no cost in believing falsely that one is well.

costs of falsely believing that *p* may be and on that basis decides how much evidence he will require before he would conclude that *p*? This sounds implausible. In particular, it sounds highly implausible to suggest that the subject would *decide* on how much evidence he will need before believing that *p*, depending on what these costs are. For surely we are all well aware of the fact that whatever costs for us may be associated with us falsely believing that *p* or not-*p* is not an *evidentially* relevant consideration with respect to that proposition. The extent to which a piece of evidence warrants a conclusion that *p* does not vary depending on what the costs are for us of falsely believing that *p*, and this is something we all know quite well. And so if we knew that we believed something partly because we lowered our requirements regarding how much evidence it will take to warrant such a belief, would this not undermine that very belief?

The idea that we could decide on how much evidence it will take for us to believe something, is not so different from the idea that we could decide to believe something, and would be just as controversial. Therefore I would say that we would be better off assuming that this is a brute causal form of explanation. Fears regarding what the costs would be of falsely believing things can have effects on our confidence thresholds, and presumably there would be good evolutionary reasons why this is so, but it is best to presume that we are looking at non-rational (though purposive) causality here. So if this phenomenon helps explain the existence of the self-deceiver's biased belief, it is a distinct kind of explanation from the more agentive explanation which I have offered, described in terms of intentional action.

However, this is not a *competing* explanation to the one in terms of biased evaluative actions, but a complimentary one. This mechanism concerning confidence thresholds may facilitate the biased evaluative actions discussed earlier in leading to a biased belief. In other words, a subject may one-sidedly seek evidence supportive of her desired conclusion, and

also have a lower confidence threshold for believing that proposition, which may incline her to hastily jump to a conclusion on the basis of those welcome considerations. But these two factors are nevertheless distinct, for there seems to be a bias in the way that the subject *acts* when thinking and searching for evidence which is *independent* of any bias or asymmetry there may be in the subject's confidence thresholds. This bias is evidenced in the greater energy they put into searching for welcome as opposed to unwelcome considerations, and again, this seems to be a different phenomenon to the bias in confidence thresholds. So we should view the FTL model as being a relevant and exacerbating factor, but not the whole explanation for how self-deceptive beliefs get generated. It is complimentary to the explanation in terms of the way the person acts.

Let me stress, then, that I don't wish to deny that the desire in self-deception may cause the belief *partly* in virtue of it playing a 'merely causal' role. However, I do think that the way it explains it by playing a motivational role is more important philosophically, since given our strong intuition that self-deception results from our actions, it is likely in virtue of the desire playing the motivational rather than the merely causal role that we are inclined to classify any given case as self-deception. In other words, if there were a case in which we had a biased belief purely due to some non-rational effect of our desire, it is not clear that this would be enough to make that a case of self-deception. Therefore, possible non-rational ways in which desires bias cognition will not be at the focus of discussion in what follows.

## *8. The Plan Ahead.*

I have put forward the suggestion that the empirical data concerning Basic Scenario cases is open to an intentionalist interpretation just as easily as a non-intentionalist one, and I have shown what these different interpretations would look like. But as yet nothing has been said

either for or against either interpretation. What, in that case, can a non-intentionalist do to motivate endorsement of his/her view of the facts over his opponent's? For one who thinks that these empirical results are open to an intentionalist interpretation, it seems that what will not be of any help in convincing him otherwise are any further appeals to empirical evidence. Rather, what would be required here is philosophical work on clarifying the concept of intentional action, and the logical implications attached to the idea of doing something intentionally. By doing this, the non-intentionalist may be able to show that an intentionalist interpretation of these facts is not a viable option after all. This will be the strategy that I will take up in chapter 5. There it will be argued that in light of certain implications associated with the idea of doing something intentionally, the intentionalist interpretation is highly implausible. The implication concerns the idea that for one to do something intentionally, one must do it knowingly, and this knowledge, it will be argued, would count as an obstacle to the success of the strategy.

Before we get involved in that, however, there is another issue that must be addressed. At the beginning of chapter 1, I said that although many philosophers hold that the Basic Scenario features A-C are not themselves sufficient for self-deception, nevertheless a lot of these philosophers would hold that paradigmatic cases of self-deception contain *at least* these features, but whether any given Basic Scenario case counts as self-deception will depend on the nature of the explanatory connection between the desire and belief (i.e., on how we answer the explanatory question for that case). Views then vary as to what this connection should be. The view I feel intuitively attracted to is that what is required to turn a Basic Scenario case into one of self-deception is that the desire that *p* motivates intentional actions which cause/sustain the unwarranted belief, and ordinary Basic Scenario cases seem to accord with this idea.

I have also offered a non-intentionalist interpretation of the intentions associated with these actions. This is the picture of self-deception I wish to defend. But some philosophers would argue that under that interpretation, these Basic Scenario cases would not count as self-deception, for the explanatory connection between desire and belief would here *not* be what is required for self-deception. What would be required, on their view, is that the desire that *p* motivate the subject to intentionally deceive herself, for this, the 'intentionality feature' as I've called it, is simply a *logically* necessary condition on self-deception. Any non-intentionalist interpretation of Basic Scenario cases would then simply exclude these cases from the category of self-deception. Until these objections are responded to, any further work will be in a precarious position. For if it is true that the intentionality feature is logically necessary for self-deception, then any attempt to show that the actions evident in Basic Scenario cases are best explained in non-intentionalist terms would only show that these cases were never ones of self-deception to begin with. The danger will be that no matter how good our explanation of these cases may turn out to be, we will never have been investigating self-deception at all.

To secure the investigation against any such misfortune, in the next chapter, I will try to defend the thesis that the phenomenon at hand, with our non-intentionalist interpretation of it, would warrant the title of self-deception, which I will do by trying to fend off various objections as to its logical insufficiency. Once that is done, we can turn, in subsequent chapters, to the issue of deciding between the intentionalist and non-intentionalist interpretation of Basic Scenario cases.

# Chapter 3: A Priori Arguments Against Non-Intentionalism

## *1. Introduction.*

In chapter 1, I presented the Basic Scenario, and mentioned a number of options for how it might be explained. In chapter 2, I showed that in real life cases meeting the conditions of the Basic Scenario, we see that the explanation is agentive. I then presented a possible non-intentionalist interpretation of the actions involved in these cases.

This phenomenon–the Basic Scenario where the explanatory connection between the desire and belief is understood in an agentive non-intentionalist way–is one that we would feel naturally inclined to call self-deception. This is a *prima facie* reason for thinking that this phenomenon *is* self-deception. Such a reason may be defeasible, and if some valid objection can be brought against it, it may be overruled. And some philosophers have objections at the ready. So the strategy here will be to defend it against a number of specific objections, and the assumption will be that if every objection can be met, then will be entitled to endorse that original intuition. The phenomenon I wish to defend as qualifying as self-deception can be captured with the following conditions, which I claim to be jointly sufficient for it. Jones is self-deceived if:

A) Jones encounters evidence warranting the belief that not-*p*.

B) Jones strongly desires that *p*.

C) The desire that *p* motivates Jones to evaluate the evidence, and seek new evidence, in a biased way, thus causing him to believe that *p*.

D) Jones' intention in doing this is just to find any evidence there might be that would confirm that *p* (he does not act with the intention to make himself acquire the belief that *p*).

(In the final chapter I will make further refinements to my set of sufficient conditions, when I recast things in terms of degrees of conviction rather than belief, but for the meanwhile we may think in terms of belief.)

The main objection to this being sufficient for self-deception that I wish to consider is that made by what I'll call *a priori intentionalism*. This is the thesis that it is a *logically necessary* condition on self-deception that self-deceivers intentionally deceive themselves. Since no such condition is mentioned above, this phenomenon wouldn't qualify as self-deception.[23]

I believe that we can identify a priori intentionalism as one of two strands of intentionalism, and I will call the other strand *explanatory intentionalism*. Explanatory intentionalists don't argue that it's logically necessary that self-deceivers intentionally deceive themselves. Rather, they argue that in cases where people believe that *p* unwarrantedly because of a desire that *p*, which they acknowledge we are inclined to call 'self-deception', the only or the best way *to explain* how this happens is to suppose that subjects make deliberate attempts to acquire that belief. Bermúdez calls this an 'inference to the best explanation' (2000: 315). I will examine and criticize this position in chapter 4. Then in chapter 5, I will show that under the assumption that the intentionalist's interpretation of the actions in Basic Scenario cases is true, these cases couldn't have obtained. In chapter 6, I will examine a version of intentionalism which holds that self-deceivers deceive themselves using a strategy which may not be subject to the difficulties discussed in chapter 5, though I

[23] Note that a priori intentionalism is quite a different kind of view from the intentionalism I spoke of beforehand. Beforehand, I spoke of intentionalism as the view that in ordinary Basic Scenario cases, people are intentionally deceiving themselves. Thus, this kind of intentionalism is committed to intentional self-deception being a real life phenomenon. However, a priori intentionalist may not be so committed. A number of people who believe that the intentionality condition is logically necessary for self-deception are skeptical about the possibility of self-deception (e.g., McLaughlin 1996).

will argue that it ultimately runs afoul of certain empirical evidence. This will then bring to an end the critique of intentionalism.

## 2. *Mele's Account of Self-Deception.*

The set of conditions I have given above is closely related to a set of conditions given by Alfred Mele which he claims are sufficient for self-deception, and which have been widely discussed. With them he intended to capture the paradigm, 'garden-variety' examples of the phenomenon. Though only meant to give sufficient conditions, let me still call this 'Mele's analysis' or 'Mele's account'. Mele has incurred objections against his account precisely on this point, however, in that it has been argued that such conditions are *not* sufficient for self-deception. Considering the similarity of Mele's account to mine, these objections would also be a threat to the account being promoted here. In what follows I will state Mele's account and give clarifications, and will then proceed to defend it against these objections. The following, then, is a statement of his conditions sufficient for being self-deceived in believing that *p*, though I have changed the wording slightly. *S* is self-deceived if:

1. The belief that *p* which *S* has is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
3. This biased treatment is a non-deviant cause of *S*'s having the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for not-*p* than for *p* (Mele 2001: 120).[24]

[24] A similar set of conditions is given by Barnes (1997: 117).

Mele's holds that one can self-deceptively *acquire* the belief that *p*, or self-deceptively *retain* the belief that *p*, and in his exact formulation of these conditions, he presents them as applying to self-deceptively acquiring the belief that *p* by, for instance, stating (1) as 'The belief that *p* which *S* acquires is false'. I have substituted 'has' for 'acquire', which gives the conditions increased generality by being ambiguous on whether the belief was self-deceptively acquired or retained.

These points are worth discussing in detail since it is the most well-known deflationist account, and so that we can see how Mele's account differs from my own, but first let me state Mele's important 'impartial-observer test' which he usually mentions immediately after stating these condition. This test is not an extra condition on top of the 4 given above. Rather, it is a test to determine whether *S*'s belief that *p* is biased and unwarranted, a point which is implicit in the above conditions. It goes as follows:

> ...given that *S* acquires a belief that *p* and *D* is the collection of relevant data readily available to *S* during the process of belief-acquisition, if *D* were made readily available to *S*'s impartial cognitive peers and they were to engage in at least as much reflection on the issue as *S* does and at least a moderate amount of reflection, those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true' (Mele 2007: 167).

It seems that in this test, the averaged judgments that the 'impartial cognitive peers' (henceforth 'ICPs') make on the basis of *D* is taken as a guide to determining the judgment that *D warrants*. If the ICPs mostly judge that not-*p* on the basis of *D*, while *S* judges that *p*, then this is good evidence that *D* warrants the judgment that not-*p* and that *S*'s judgment that *p* is therefore unwarranted and also biased in favour of what she wants to be true.

Furthermore, if the only relevant difference between *S* and her ICPs is that *S* desired that *p* while the ICPs lacked any strong preference regarding whether *p*, then the deviation in *S*'s judgment from what's warranted can be put down to her having that desire. On Mele's account, assuming that *p* is false and that the causal chain between *S*'s desire and belief is non-deviant, this means that all four conditions above have been satisfied and *S* is self-deceived. This test will be discussed further at a later point.

### *3. Further Discussion of the Elements of Mele's Account.*

This section will be spent clarifying Mele's sufficient conditions individually. Though this is useful to help us understand this analysis, it will also help us to see how closely related Mele's analysis is to my interpretation of the Basic Scenario.

(1): Regarding the first point, Mele thinks that for purely 'lexical' reasons, the self-deceptive belief must be false. He says, '[a] person is, by definition, *deceived in* believing that *p* only if *p* is false; the same is true of being *self-deceived in* believing that *p*' (1997a: 95). This point seems to be accurate, as far as it goes. It seems that one can only be deceived in believing that *p* if *p* is false. Being deceived in believing that *p* is in this respect is like being *mistaken in* believing that *p*. However, there is a certain point that should be noted here. Mele fails to recognize that there is another grammatical construction for attributing self-deception that does not carry this false belief entailment. This construction is of the form '*S* deceived herself into believing that *p*'. Being deceived *in* believing that *p*, and having been deceived *into* believing that *p*, should be distinguished, and I would suggest that where it is contradictory to say that *S* was self-deceived in believing something true, as Mele claims, it is

not contradictory to say that *S* deceived herself into believing something true (on this point, also see McLaughlin 1988: 35-36. Barnes 1997: 8).

Support for this idea can be found by looking at cases of interpersonal deception. Say that I know that *p* is true and I know you will only believe that *p* if you hear it said from Jones, who is the only man you trust on this matter. I then lie to you by saying that I was in touch with Jones, who told me that *p*, and you then believe that *p*. In this situation, you are not *deceived in* believing that *p*, since this is a true belief. However, we can say here that you were *deceived into* believing this true proposition by me. This is like saying you were *tricked into* believing that *p*, which one can be even where *p* is true. Though *p* was true, I utilized deceptive means in order to get you to believe it. So condition 1 should not be regarded as a strictly necessary condition for self-deception. For these reasons, I do not have an equivalent condition in my own account.

(2) & (3): Perhaps the above quibble is a relatively pedantic matter, and not much of significance hangs on it. The next few points are more important. Condition 2 states that '*S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.' What does he mean by 'motivationally biased way'? In an earlier formulation of condition 2 from a 1983 article, Mele phrases it as, '*S*'s desiring that *p* leads *S* to manipulate (i.e., to treat inappropriately) a datum or data relevant, or at least seemingly relevant, to the truth value of *p*' (1983: 370). It seems, however, that he came to want to expand his account to include cases of 'twisted self-deception', and he reformulated it accordingly. In these 'twisted' cases, as Mele sees them, the subject also believes that *p* unwarrantedly, but this is caused by his strong desire *that not-p* (or his aversion/fear that *p*) 'biasing his treatment of the

data'[25]. So by '*S* treats data…in a motivationally biased way' Mele means that *S* either desires that *p* or that not-*p*, and this leads him to treat the data relevant to *p* in a biased way. In this thesis we are limiting our concern to cases where the subject believes that *p* because of a desire that *p*, so it will be better to work with Mele's older formulation of condition 2. This formulation also states the condition in terms of desire rather than the semi-technical term 'motivation', and may be more perspicuous for that reason. The set of conditions with 2 formulated in the old way is:

1. The belief that *p* which *S* has is false.
2. *S*'s desiring that *p* leads *S* to manipulate (i.e., to treat inappropriately) a datum or data relevant, or at least seemingly relevant, to the truth value of *p*.
3. This biased treatment is a nondeviant cause of *S*'s having the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for not-*p* than for *p*.

These conditions Mele would regard as being sufficient for self-deception in its *straight* variety. We may regard the desire here, with Barnes (1997), as an *anxious* desire, which may be the only kind of desire powerful enough to give rise to such distortions in the reasoning process.

Points (2) and (3) here can be regarded together as parts of a single causal process. That is, *S* desires that *p*, and ends up having the belief that *p*, where the desire that *p* figures in the causal story of how *S* acquired or maintained that belief. Moreover, I find it to be a natural reading of these conditions to interpret this causal process as a motivational one, in

[25] Note that this is not the only way to understand these cases. According to Barnes (1997), in these cases what is causing the biasing, leading to the holding of an unwelcome, unwarranted belief that *p*, is another desire, the desire that *q*, which the belief that *p* functions to placate.

*my* sense of 'motivational', for Mele speaks of the desire as leading *S* to *manipulate* or *treat the data* in a biased way, where that treatment causes/sustains the unwarranted belief. These sound like actions on the part of *S*, actions motivated by the desire that *p*, and so this allows us to give Mele's account an *agentive* reading, despite Mele's sporadic remarks here and there suggesting that he is a non-agentive deflationist.

For Mele, the desire must also cause or sustain the belief in an appropriate way: a 'non-deviant' way. As Mele puts it in the earlier formulation of this condition, there must not be a deviant causal chain 'between the desire and the manipulation or…between manipulation and belief-acquisition' (1983: 370). What would a 'non-deviant causal chain' look like? How one answers this question may depend on one's general theoretical persuasions, and in effect, this question will be the ongoing concern of this chapter. A priori intentionalists, for instance, would claim that a causal chain involving an intentional act of self-deception would be a non-deviant causal chain, and anything less than that would be deviant (and this is the position I wish to criticize in this chapter). Mele wouldn't agree, though he might have a more liberal view of what a non-deviant causal chain would be than I would like to accept, which would tolerate an explanation purely in terms of non-motivational (in my sense of 'motivational'), 'merely causal' factors or processes. In the view being advocated here, the standard non-deviant causal chain would be one where the desire that *p* motivates *S* to 'treat the data', that is to say, reason about, evaluate, or search for data, in a biased way (as described in chapter 2), thus causing himself to have the unwarranted belief (facilitated perhaps by biases in his confidence thresholds), where these actions were not done as part of an attempt to intentionally make herself have that belief.

The following, on the other hand, would be something we could all agree on as being an example of a deviant causal chain. Suppose a brain scientist put an implant in *S*'s brain, such that when *S* desires that *p*, and then manipulates the data is a biased way, this

automatically triggers a mechanism in the implant which causes *S* to acquire the belief that *p* (this fantasy might be unintelligible when thought through, but waive that for the moment). This would be a situation where the desire causes the belief, but probably no one would regard this as self-deception. In my view, all causal chains from the desire to the belief which are *non-motivational* are deviant.

(4): Now for condition 4. Mele claims that condition 4 is expendable. Some cases of self-deception may not involve *S* possessing evidence that warrants overall the belief that not-*p*, and he seems to just add it in for good measure. The cases Mele has in mind here are ones in which the evidence readily *available* to *S* warrants believing that not-*p*, though he is not in *possession* of (acquainted with) that evidence. However, because *S* goes about gathering evidence in a motivationally biased way, he ends up being acquainted with a biased stock of the available evidence, from which he concludes that *p*, whereas had *S* gathered evidence in an unbiased way he would have concluded that not-*p*.

However, point 4 may be more important than Mele realizes, and in fact, the idea that 4 is not necessary is not in harmony with another thesis which he shows a measure of commitment to. I have in mind his proposal for how self-deception and wishful thinking are to be distinguished. Mele rejects the claim, sometimes advanced by intentionalists, that wishful thinking is distinguished from self-deception through the latter being brought about intentionally while the former is not. He then proposes a way of distinguishing them, one that, incidentally, is the favoured view among non-intentionalists: 'the difference may lie in the relative strength of relevant evidence against the believed proposition: wishful thinkers may encounter weaker counterevidence than self-deceivers' (1997a: 100). Note his use of the word 'encounter' here, which means that self-deceivers are acquainted with or in possession

of this relatively strong counterevidence, and not that it is merely available for or obtainable by them.

So on this view (one which I will defend later), self-deceivers acquire their belief against or 'in the teeth of' relatively strong or significant counterevidence, which distinguishes them from wishful thinkers, who believe unwarrantedly but not in the face of significant counterevidence. This suggests that unless condition 4 is also met in a given case, it may with more right be called wishful thinking, by Mele's own criteria. So without condition 4, Mele may have trouble respecting the distinction between self-deception and its (perhaps) closest relative. Indeed, some philosophers who note Mele's comment that point 4 is not necessary have found reason to complain with the overall analysis on that basis, on the grounds that it conflates self-deception with wishful thinking (Scott-Kakures 2002). Later the distinction between wishful thinking and self-deception will be discussed in more depth, and the position will be taken that for the distinction between the two to be respected, point 4 should be counted as necessary.

It may be observed that Mele's analysis pretty much captures the same phenomenon that I have been trying to capture with the non-intentionalist interpretation of the Basic Scenario cases in chapter 2, and any objections as to its sufficiency for self-deception will be objections to the current position. We will now turn to these objections. Many of them tend to be or may be put forward by what I will call *a priori intentionalists*. A priori intentionalists consider it to be a *logically* necessary condition on self-deception that it be brought about intentionally (what we may call 'a priori intentionalism'), and would cite the omission of this condition in Mele's account as its shortcoming. I will divide the arguments for this view into two categories. The first category represents a number of arguments that we may call *distinguishability arguments*. These include arguments that have been made in the literature,

and arguments that I anticipate could be made, and they argue somewhat indirectly for the necessity of the intentionality condition. They work by trying to show that the above conditions don't succeed in distinguishing self-deception from some other kindred phenomenon which it is evidently distinct from, and so do not give sufficient conditions for self-deception as the deflationist claims. It may then be suggested that the intentionality condition is the best candidate for distinguishing it from this closely related phenomenon.

After that, I will consider what I call the *lexical argument* for a priori intentionalism, which is a very important argument in the history of the debate. According to this, this deflationary analysis fails because it strays too far from the model of interpersonal deception, and self-deception should be understood on that model. Modeling self-deception on interpersonal deception, we get the result that self-deception must be intentional (and according to some, we also get the result that it must involve the contradictory beliefs condition). I will defend the sufficiency of the deflationary analysis against these objections in what follows. Addressing these arguments also offers us the chance to increase our understanding of self-deception by situating it in the landscape of related psychological phenomena, and with respect to the other species of deception.

### *4. Distinguishability Arguments Against Non-Intentionalism.*

*Mistaken belief*. According to one distinguishability argument, if when a self-deceiver deceives himself into believing the false proposition *p*, he does not do this intentionally, then he must do it unintentionally. Therefore, it would amount to a case of making an unintentional mistake or error in one's reasoning or assessment. But this steamrolls over an important distinction. Susan might divide *x* by *y* and get *z*, which is the wrong answer. Here she will also have come to hold a false belief, unintentionally, due to her own activity and

through her own fault. But this would hardly be self-deception. Self-deception must be distinct from such cases. But what distinguishes it? One might then go on to think that a compelling answer to what distinguishes self-deception from cases of this sort is that in self-deception the mistaken belief was intentionally brought about (Steffen 1986: 45).

This argument is easily disposed of. According to non-intentionalism of the kind being promoted here, self-deception does not *just* involve unintentionally making oneself come to have a mistaken view. It involves that, but where the explanation for how one made oneself have that mistaken view has to do with one having a desire that that view be true, or a fear that it be not true: to 'motivational' influences on belief. The idea that Susan *had a stake* in $z$ being the correct answer, and had a pre-existing *desire* for $z$ to be the correct answer, is not part of the explanation for why she mistakenly came to believe that $x$ multiplied by $y = z$, and that's why she was not self-deceived in so believing. The motivational element can do the work of distinguishing being self-deceived in believing that $p$, from other cases where we make a mistake in our reasoning and thus unintentionally make ourselves come to believe something erroneously.

*Prejudicial or racist belief*. Richard Holton has offered the following more serious case as a counterexample to Mele's conditions, describing it as follows:

> Jean-Marie is a racist. He thinks that blacks and Arabs are not as good as whites: not as clever, or as imaginative, or as brave, or as trustworthy, or whatever. Take just about any property that Jean-Marie might regard as a virtue, and he will think that whites have more of it. Let us assume that his beliefs here are, by and large, false. But he holds them sincerely. And were we to challenge them, he would provide evidence: reams of it, taken from the magazines and newspapers of the kinds of organisation to which he belongs. He is aware of the opposing view; indeed he has reams of

> [evidence supporting the other side] too, collected to document the conspiracy which he thinks pervades the liberal establishment that controls the mainstream press and publishing houses (Holton 2001: 59).

Holton supposes that in this case Jean-Marie meets all Mele's conditions, and yet we would not, he claims, regard him as self-deceived, but just bigoted and prejudiced. Holton does not use this case to argue for intentionalism himself, but one could easily imagine someone claiming that the inclusion of the intentionality feature would be what's needed to turn this case into a genuine one of self-deception.

The problem with this example is that it is not made evident from Holton's description of the case that it satisfies all of Mele's conditions, and if that were made clear, it is not so evident that we would remain as disinclined to think of it as a case of self-deception as Holton assumes (for saying that a belief is racist does *not* preclude that it is *also* self-deceptive). Included in Mele's conditions is the point that the subject treats the data (i.e. searches for and reasons about the data evidentially relevant to the issue) in a biased way, where this biased thinking and searching be caused or motivated *by the desire that p*, which in turn causes her to have the belief. And although there is a suggestion that Jean-Marie has engaged in biased reasoning and evidence searching about the issue, from the description of the case given we are not told that Jean-Marie anxiously *desires* that blacks and Arabs be inferior to whites, and that this is what's driving the biased behaviour. And if this was made explicit, a diagnosis of self-deception would begin to look much more plausible.

Generally speaking, it is not a conceptually central feature of prejudicial or racist beliefs that the subject has a strong desire for the believed proposition to be true. The existence of a prejudicial belief may not even be the outcome of any biased reasoning as such, and may often be explained with reference to various non-motivational influences, such

as cultural influences. Such beliefs, and the erroneous premises on which they are based, may get ingrained in one through the influence of associates, of education, customs, or through features of the social structure. They may also arise from indoctrination. In such cases, a desire that the proposition be true need not feature in the explanation of the prejudicial belief at all, and furthermore, neither may the belief have resulted from biased reasoning (the belief may simply have been adopted from these influences and authorities, without much reasoning having taken place). On the other hand, such a desire *is* a central feature in the explanation of *self-deceptive* belief. So if it were true that the racist's belief that blacks are inferior was caused or sustained because of his desiring that they be inferior (perhaps because he has a personal stake in a social system legitimized through that assumption, as did the white beneficiaries of Apartheid or slavery), where this desire drove him to reason in a biased way about the issue, then self-deception would be a better diagnosis of the situation, and indeed, it is quite plausible that in many instances, racist beliefs *are* sustained by self-deception.

*Stubborn belief*. Another kind of belief that is akin to self-deceptive belief is stubborn belief. Stubborn belief may also be motivated or desire-influenced, and so one might wonder whether the 4 conditions given above give us the resources to distinguish self-deceptive belief from this phenomenon, or whether an added intentionality condition is needed. I believe that they do allow us to distinguish the two. In typical cases where *S* is being stubborn in holding to the belief that *p*, this will indeed be due to desire biasing his reasoning. But it will not be the desire that *p* be true in particular. Rather, it will be something like the desire to be right, or to not to have to revise one's beliefs, or to not have to go to the effort of inquiring further into the matter, or not to be shown up as being wrong by some irritating or dislikeable person. Stubborn believers, unlike self-deceivers, don't seem to believe as they do because they have a particular stake in whether or not the state of affairs, *p*, is the case, and they would equally refuse to listen to reason if it were some other random proposition that was

under consideration. For they may just hate being proved wrong or hate having to change their minds, and refuse to listen to criticism or to reconsider the matter because of that.[26]

*Wishful thinking*. In the above cases, the fact that the unwarranted self-deceptive belief is caused by the desire that *p* could be seen to distinguish such beliefs from beliefs of the other kinds. However, we cannot appeal to this feature to distinguish self-deceptive belief from belief born of wishful thinking, since wishful thinking, as is generally maintained, does not differ from it in that respect (Gardner 1969-70: 242). Here we are perhaps faced with the most serious of the distinguishability arguments. According to this one, wishful thinking involves coming to believing something unwarrantedly because you have a stake in it and want it to be true, just as the non-intentionalist says self-deception does. What, then, distinguishes self-deception from wishful thinking? An intentionalist can propose that although both self-deception and wishful thinking involve believing that *p* unwarrantedly because you desire that *p*, it is the intentionality of the former that distinguishes it from the latter. That is, in self-deception, the desire that *p* causes the belief that *p* by leading to an intention to make oneself believe that *p*, whereas in wishful thinking, the desire that *p* causes the belief in some other way (e.g. Bermúdez 2000: 312). The burden then falls upon the non-intentionalist to find a plausible way of distinguishing between wishful-thinking and self-deception without appeal to such a feature.

Indeed, non-intentionalist philosophers have proposed a way of distinguishing wishful thinking from self-deception under the presumption that in both one believes that *p* because of a desire that *p*, without appeal to an intentionality condition. And not only have they presented a different way, but that different way is a plausible and intuitively satisfying one, or so I shall argue. Because critics of deflationism quite often object that it can't distinguish self-deception from wishful thinking, and because wishful thinking is possibly self-

[26] There may be a *loose* sense in which stubborn belief might also be called self-deception, but in the primary sense, self-deception can, I think, be distinguished from stubborn belief along these lines.

deception' closest cousin, so to speak, this issue will be given a fairly sustained discussion in the next section.

## 5. *Wishful Thinking and Self-Deception.*

To get an initial grip on the phenomenon of wishful thinking, let's consider what might be paradigm cases of it. Consider the phenomenon of *unrealistic optimism*. In one study, a large group of undergraduate students had to indicate on a questionnaire how likely they were to suffer in the future from a range of different health ailments relative to the average person in the group. Most subjects believed themselves to be less likely than the average person to experience them (Weinstein 1982). Similarly, when another group of students were instructed to indicate on a questionnaire how likely they were to experience a range of events in their future, events which were either positive (e.g. 'weight constant for 10 years', or 'like postgraduation job') or negative (e.g. 'having a drinking problem' or 'divorces a few years after marriage'), the majority of subjects tended to think they were more likely than the average person in the group to experience the positive events, and less likely than the average person to experience the negative events (Weinstein 1980). In another, similar study, subjects were asked how they rated themselves in terms of skill and safety at driving relative to other drivers. The majority indicated that they were better than average on both dimensions (Svenson 1981). Other surveys have shown that the average person has a tendency to think that he/she is more intelligent, more attractive, and more popular than the average person (Balcetis 2008). These studies seem to exemplify the phenomenon of wishful thinking.

These phenomena look quite similar to the My Lai Massacre cases which I offered as exemplars of self-deception, but there are important differences. In the latter case, people encountered considerations which strongly suggested that an unwelcome proposition was

true, and yet they resisted the thrust of this evidence, taking up a hypercritical stance towards it. There was a *defensive reaction* to unwelcome evidence, an attempt to explain it away. However, in the cases we are now looking at, a lot of the people who think they are above average on these dimensions must realistically have been around average relative to their peers. So presumably they had no strong basis either for or against the proposition that they would like to be true: that they are above average. So it looks as though most of these cases may not have involved subjects *resisting evidence* or taking a *defensive reaction to unwelcome evidence* that supports an unwelcome proposition. According to Weinstein, one of the ways in which these biased judgments occur is by the subject bringing to mind personal attributes and experiences that suggest that they are well positioned on the relevant dimension, neglecting to appreciate that the others may be able to produce similar evidence for themselves (1980: 819). So these cases may involve no activity of taking up an attitude of skepticism towards, or of explaining away, unwelcome evidence.

A related kind of case is illustrated in an experiment by Ditto *et al* (1998). This case shows how people can make an irrational judgment in the direction of what they want to be true in response to a modicum of evidence that is *supportive* of that judgment, which they prematurely accept too uncritically. In an experiment ostensibly on 'the accuracy of first impressions', subjects who were all males filled out a personality questionnaire that was then given to an attractive woman (a confederate) who was supposedly to form an impression of them on its basis. Her written evaluation was then given to each subject for them to read, and was either flattering or unflattering. Subjects were led to believe either that she had been free to write both what she liked and disliked about him, or that she had been instructed to focus specifically on what she liked, or in other cases, disliked about him. Afterwards, on a questionnaire about the woman's evaluation, each subject had to indicate how much he thought she liked him.

It was found that subjects who received a flattering evaluation, but who were told that the woman was constrained to write only what she liked about him, attributed to her overall impressions of them that were as positive as the impressions attributed to her by subjects who received flattering reports but were told she was free to write whatever she wanted. In other words, these former subjects were shown to have failed to adequately take into account the background context of her evaluation. They credulously accepted the positive evaluation as indicating a positive overall impression of them, and were insensitive to the constrained nature of that evaluation. On the other hand, subjects who received an unflattering evaluation were sensitive to the background context. When they knew the evaluation was unconstrained, they attributed to her a poor opinion of them. And those who were told she had to focus on the negative adjusted for this, and took her impression to be less negative, indicating something more like uncertainty regarding her opinion. Although in this case the judgments of the people who received the flattering evaluations under the constrained conditions were not compared to the judgments of any disinterested observers asked to judge the woman's impression of these male subjects, it is plausible to presume that such a group of onlookers would have taken account of the situational context and not judged her impression of them to be so positive.

Three kinds of case can now be distinguished, which seem to occupy positions along a continuum, ranging from those who are presented with unwelcome evidence, to those lacking significant evidence either way, to those presented with some welcome evidence. We can call them cases (A), (B), and (C).

(A) There are cases where people judge unwarrantedly in the direction of what they would like to be true against evidence to the contrary, because they want it to be true.

(B) There are cases where people judge unwarrantedly in the direction of what they would like to be true when they have no significant evidence either going against or going for that position, because they want it to be true.

(C) There are cases where people judge unwarrantedly in the direction of what they would like to be true after receiving a modicum of evidence pointing in that direction, because they want it to be true.

It seems to me that we are inclined to categorize both B and C-type cases as wishful thinking. For instance, the Weinstein study, where people were shown to have unrealistically optimistic views about their own future, is as stereotypical a case of wishful-thinking as you could hope for, and the subjects in the Ditto *et al* (1998) study also strike us as wishful thinkers. However, it should be noted that we are more inclined to categorize A-type cases differently, namely, as cases of self-deception. For instance, the My Lai Massacre cases seemed like typical cases of self-deception, and these are A-type cases.

This suggests the following account of the distinction between self-deception and wishful thinking. What distinguishes the former from the latter is that self-deceivers, essentially, *are faced with evidence which goes against* what they want to be true, while wishful thinkers are not. Because self-deceivers have to hold a belief in the face of such evidence, self-deception essentially involves taking *a defensive reaction against unwelcome evidence*, or *resisting the thrust of the evidence*, unlike wishful thinking. This resistance and defensive reaction will require classic behaviours associated with self-deception, including explaining away or reinterpreting evidence in innocuous ways, behaviours which nobody associates with wishful thinking. Note, however, that *no* mention need be made of any intentionality condition in distinguishing these phenomena here.

Indeed, quite a number of philosophers have described the distinction between self-deception and wishful-thinking along these lines (see Bird 1994: 37-38. Cosentino 1980: 455. Graham 1986: 224-225. Mele 1997a: 100. Szabados 1973, 1974: 60-61.). For instance, in his excellent paper on the distinction, Béla Szabados speaks of a man in love with a woman who sees reciprocal feelings behind her acts of friendliness towards him. Szabados claims that he is guilty of wishful-thinking insofar as he 'jumps to conclusions from shreds of evidence motivated by his desire to be loved by her', while he passes into self-deception insofar as he is presented with significant evidence that she's not (she's been seen having intimate moments with another man) and then 'proceeds to resist, by ingenious tactics, the natural implications of the evidence' (Szabados 1973: 204-205). Again, we need not advert to any intentionality condition to plausibly account for the distinction here.

Admittedly, the expression 'self-deception', is sometimes used quite loosely in ways not always entirely sensitive to the subtle yet important distinctions between A-type cases, and B and C-type cases. In the psychology literature, for instance, people sometimes do discuss B and C-type cases under the heading of self-deception. Granting that many B and C-type cases seem like paradigms of wishful thinking, this would suggest that wishful thinking is a species of self-deception, and not a distinct phenomenon from it. What are we to make of these vagaries of use? One view voiced by some philosophers is that perhaps the concept of self-deception is not as sharp or determinate as a psychological taxonomist would hope for (see McLaughlin 1996: 45-46. Audi also suggests the term 'self-deception' has 'no settled use', 1976: 381). In that case, our account of the distinction between self-deception and wishful thinking could be interpreted as at least partly prescriptive, and not just purely descriptive, of our pre-theoretical word usage, an attempt at sharpening a concept whose application to these cases is not already prescribed by entrenched conventions. I do not share this pessimism however. My own view is that these are simply careless uses of the word, and

that philosophical reflection, done by comparing cases and noting their distinguishing features, exposes them as such for anyone who so cares to reflect.

In summary, our deflationary account of self-deception gives us the resources to distinguish self-deception from all the above kindred phenomena without needing to appeal to the idea of intentions to deceive oneself. And though there may be other kindred phenomena we haven't thought of here, it's likely that it would allow us to distinguish self-deception from those phenomena too, seeing as it does seem to allow us to distinguish self-deception from these which seem like its *closest* psychological relatives.

## *6. The Lexical Approach to the Analysis of Self-Deception.*

How similar self-deception has to be to interpersonal deception for it to remain a species of deception is an issue that has come up time and again in the self-deception debate, and one of the main criticisms of non-intentionalist approaches such as Mele's is that they stray too far from the model of interpersonal deception, thereby being guilty of an undue 'stretching of the term'. The intuition that self-deception ought to be understood on this model has proven to be tenacious, and it usually leads people to think that self-deception must be intentional. It also sometimes inclines people to think that believing that *p* and that not-*p* simultaneously must be a feature of the self-deceived condition. A non-intentionalist who wants to defend his/her thesis against this common thought may be obliged to identify some erroneous presupposition that is generating it. This is the strategy that will be taken in what follows.

The approach that analyses self-deception on the model of interpersonal deception has been called by Mele the 'lexical approach', and he contrasts this with the methodology that he favours. On the lexical approach, we 'start by asking what "deception" (or "deceive") means, and then ask what "self-deception" must mean if it is to be a species of deception'

(1987: 13). That is to say, we derive the meaning of 'self-deception' from the meaning of 'deception' as it is used to refer to the interpersonal case. This contrasts with what Mele calls the 'example-based' approach. On this approach, '[o]ne starts by gathering and constructing cases that would generally be described as self-deception, and then attempts to develop an analysis of self-deception on the basis of a consideration of this material. The meaning of "self-deception" is determined by the cases, which are therefore the most fundamental data' (1987: 13-14).[27] This approach is therefore a more *direct* approach, which does not try to reach an understanding of self-deception *via* the understanding of interpersonal deception. It advocates just looking towards paradigm cases of self-deception without considering cases of interpersonal deception at all. Shortly we will see how it is that both approaches tend to yield very different answers to the question of what self-deception is.

I think that we can understand the lexical approach as being guided, whether explicitly (see Gozzano 1999: 141) or implicitly, by the following formula. We can call it the *lexical formula*:

> What it means for someone to deceive himself is for him to do the same thing to himself that he does to another when he deceives another.

To many an ear this formula may sound intuitively compelling. And perhaps one might try to justify it with reference to other reflexive expressions. For example, if Jones shoots Smith, Jones points a loaded gun at Smith and pulls the trigger. And what else, in that case, could it be for Jones to shoot himself, if not to do *that very same thing*, but to himself, namely, point a

[27] After making this two-fold distinction in a 1987 paper, he later introduced another category: the 'theory-guided approach'. As he defines it, on this approach 'the search for a definition is guided by commonsense theory about the etiology and nature of self-deception (1997: 92). However, Mele doesn't say anything about what these 'commonsense theories' are, and it is unclear what philosophical accounts he has in mind as exemplifying this approach.

loaded gun at himself and pull the trigger? And if he were to do anything other than just that, wouldn't it be doing something other than shooting himself? Pataki displays his attraction to this idea when he says that '[m]ost of the reflexive attitudes and actions, for example, self-appraisal, self-loathing, self-aggrandisement, retain the essentials of their non-reflexive prototypes and it would be surprising if self-deception were an exception' (Pataki 1997: 322).

On the lexical approach, before we can employ this formula to see what it is to deceive oneself, we must first establish what it is to deceive another. The following definition is usually taken to capture what this is (e.g. Gardiner 1969-70: 224. Bird 1994: 20):

> When A deceives B, A intentionally causes B to believe something that A knows/suspects is false.

It is interesting to note that many philosophers are just as inclined to use the term 'deliberately' as they are 'intentionally' when talking about deception, treating them as practically interchangeable in this context (see Bird 1994: 20. Davidson 2004/1986: 206. Haight 1985: 245. Rorty 1988: 12. Bach 2009: 782), and perhaps some would accept that deliberateness is just as central a notion as intentionality in interpersonal deception. An interesting question would be whether they can come apart, that is, whether it would be possible for A to deceive B deliberately but not intentionally, or intentionally but not deliberately, but I will not pursue this taxing topic here and will stay with the notion of intentionality (for a discussion of the distinction between these two notions, see Austin 1966).

So on the above, commonly encountered definition of interpersonal deception, a salient feature is that the deceiver intentionally commits the deception. Where A accidentally misleads B into believing that *p*, or makes B believe *p* without knowing that *p* is false, we typically withhold the charge that A *deceived* B, which has more severe undertones (see

Pataki 1997: 311). In this way it would be thought that the act of deception is one that necessarily must be done intentionally. Perhaps we should not find this so surprising, after all, there does seem to be a host of activities that can only be carried out intentionally: it's not obvious how one could unintentionally lie, steal, murder, apologize, plot, or seduce. Could deception just be of a kind with these?

One does not find very much by way of arguments or justifications for this definition in the writings of philosophers who work with it. It tends to be adopted quite casually for the most part, as if it were obvious. I do not want to suggest that it is beyond reproach. Perhaps Mele is right in saying that the word 'deceive' may sometimes be used to refer to cases of unintentional misleading (1997a: 92), though I would imagine this is a rather loose use of the term. And it is true that some clever counterexamples have been advanced against this definition, though these tend to make reference to odd, atypical kinds of case (e.g. Barnes 1997: 8-11. Champlin 1977: 289). However, I do not want to focus on any problems there may be with this stage of the lexical argument, since I believe the most serious problems with it lie elsewhere. I am therefore willing to take it for granted from here on in. What perhaps could be said in favour of this definition is that it probably is successful in capturing the central or stereotypical cases of interpersonal deception, which is something that even critics of this definition such as Mele tend to acknowledge (1997a: 92). In this way perhaps traditionalists would have scope to argue that the counterexamples are conceptually peripheral or limiting cases. At any rate, let's take it for granted for now. Then, feeding this definition into our formula, we can easily deduce a priori that:

> When A deceives himself, A intentionally causes himself to believe something he knows/suspects is false.

Now if we assume the validity of the lexical formula, and of the definition of deception derived from the interpersonal paradigms, then it follows logically that self-deception must be as this definition says. On this picture of self-deception, A, after encountering evidence that makes him realize that some proposition *p* is true, intentionally causes himself to believe the contrary proposition, not-*p* (usually for the sake of avoiding the anxiety associated with knowing that *p*).

Some 'lexicalists' also think that we get the result that A ends up in a condition where he believes that *p* and believes that not-*p* simultaneously. How the approach leads to this conclusion is usually explained, in a rather glib fashion, along the following lines: if in the interpersonal case, the deceiver must believe (the truth) that not-*p* while the deceiver must end up, because of the deceiver's actions, with the belief that *p*, then in the reflexive case, where the deceiver and the deceived are one and the same person, this person must in his capacity as deceiver know that not-*p* and also in his capacity as deceived believe that *p*.[28] However, even if we assume that self-deception should mirror interpersonal deception in this way, the point begs the question of why the subject must satisfy the role of deceiver and deceived simultaneously rather than consecutively. For although it is typical of interpersonal deception that the deceived, B, believes that *p while* the deceiver, A, knows that not-*p*, this is not necessary for interpersonal deception (A could send B a deceptive letter and die or be convinced otherwise while it's in transit, see (Siegler 1963: 35)). So might it not be the case that the self-deceiver starts out knowing that not-*p*, does something to make himself believe that *p*, and by doing so he 'exchanges' one belief for the other?

David Kipp criticizes this idea by asking rhetorically 'is the supposed process of change from mere deceiver at one time to mere deceived at some later time conceivable apart

[28] For instance, Sartre says 'the one to whom the lie is told and the one who lies are one and the same person, which means that I must know in my capacity as deceiver the truth which is hidden from me in my capacity as the one deceived' (Sartre 1956).

from assuming a transitional stage in which one is simultaneously both deceiver and deceived? If not, as I would hold, we are returned to the very paradox that successivist interpretations…are intended to evade' (Kipp 1980: 309). But we no more need to suppose such a transitional stage between the conditions of believing exclusively that not-*p* to believing exclusively that *p*, where the subject believes that *p* and that not-*p* simultaneously, then we need to suppose an equivalent transitional stage for someone who kills herself, in between the stages where she is alive and then dead, where she is both alive and dead simultaneously. Kipp gives no argument for this transitional stage thesis, and there are clear-cut cases of intentional self-deception involving a transition from believing that not-*p* to believing that *p* using a strategy involving the exploitation of one's poor memory, which involve no such transitional stage[29]**.** Thus, it seems that there is little reason to think that from applying the lexical formula to the above definition of deception, the contradictory belief feature follows logically.

As I remarked earlier, the traditionalist approach to self-deception is associated with two commitments: that the self-deceived condition involves the subject holding contradictory beliefs, and that it is brought about intentionally by the subject. The lexical approach, then, does not hold out great promise for a proof that the very concept of self-deception entails the first feature (later we will address another more serious argument for this feature, which is an explanatory argument[30] rather than an a priori one). Nevertheless, on the whole it seems that if we grant the lexical approach, then there is a reasonably strong case to be made for the

[29] The case often mentioned in the literature is what I'll call The Diary Case. In this case, a woman has an appointment at some future date which she doesn't want to attend. She then intentionally attempts to fool herself into thinking that it is on a date she knows it's not on by putting a note in her diary that it's taking place on that date, knowing that she will probably have forgotten her deceptive plan when the time comes and so will believe that the date is correct when he sees it (see McLaughlin 1988: 31-32. Mele 1997: 99. Bok 1980: 928). A similar but more dramatic example can be seen in the movie *Memento*, directed by Christopher Nolan.

[30] That is to say, it is an argument to the effect that we need to suppose that self-deceivers have contradictory beliefs in order to explain a certain feature of self-deception.

second feature (that self-deception must necessarily be intentional), which would vindicate a large portion of the traditionalist's view. Note that it wouldn't yet follow from all this that self-deception actually exists. On the lexical approach, the philosopher asks *hypothetically*: 'What conditions would *have to* be met for there to be a case of self-deception?' For that reason, there is room left for the question of whether self-deception ever obtains at all or whether it is even possible, and as Mele points out, many who take the lexical approach end up being skeptics about self-deception (1997a: 92).[31]

It is also worth mentioning that although, as I have argued, the notion of contradictory beliefs is not an essential component of the notion of intentional self-deception (as the diary case shows, for instance, see note 30), later, when we work out in more detail what most intentionalists think intentionally deceiving oneself actually involves in practice, we will see that it does involve putting oneself into a state where one has contradictory beliefs. But this is different from saying that the contradictory belief feature is entailed by the idea of intentional self-deception. The more immediate task, however, is to examine the validity of the lexical approach to the analysis of self-deception itself.

## *7. Problems with the Lexical Approach.*

There are two ways that one could undermine this lexicalist reasoning for a priori intentionalism. Firstly, one could argue that the definition of interpersonal deception is incorrect.[32] In particular, one might try to argue that interpersonal deception doesn't have to

[31] Note that this skepticism is ruled out from the outset by the example-based approach, since this takes the legitimacy of people's ordinary use of 'self-deception' for granted and just asks what goes on in the actually existing cases so referred to.

[32] The definition of deception as involving A making B believing something he knows is false may be overly-simplistic, and more nuanced accounts of interpersonal deception countenance other possibilities. Chisholm and Feehan, for instance, think that A deceiving B can also involve A causally contributing towards B continuing in

be intentional. This is the strategy that Mele takes, who tries to make something of the fact that 'deceive' is sometimes used without any implication of intentional deception taking place (he notes, for instance, that one might say 'unless I am deceived, I left my keys in my car' (2001: 8)). I do not find this strategy particularly satisfying, since it tries to derive a substantive philosophical conclusion about the concept of deception from uses of the word 'deceive' which seem like figures of speech, or loose and peripheral to its primary, central use where it seems to carry the implication of intentionality (this Mele somewhat admits in granting that such cases are not *stereotypical* case of deception). The other option is to try to undermine the validity of the lexical formula. This is the strategy I will take as it seems to me that the deepest problems lie with this. So at least for argument's sake, let's grant the intentionalist the definition of interpersonal deception, and turn to the lexical formula.

If the lexical formula is valid, then we should expect it to give us the right result in cases other than that of deception. We should expect that we could turn it into a general formula for deriving the meaning of any reflexive construction grammatically analogous to 'deceiving yourself', i.e. expressions of the form '*V*ing yourself', where '*V*' represents some verb. Accordingly, we can state this generalized version of the lexical formula as follows:

> What it means for one to *V* oneself is for one to do the same thing to oneself that one does to another when one *V*s another.

…where *V* stands for some verb. But as T.S Champlin has argued (1977: 284-285. 1988: 24-25, also see Martin 1986: 20), such an approach can be parodied for a number of such

---

a false belief, or towards B losing a true belief, or may involve A preventing B from acquiring a true belief (1977: 144). I doubt whether merely withholding information in sufficient for deception however. Chisholm and Feehan do not discuss the point of whether these actions would have to be carried out intentionally.

reflexive constructions. Consider how the lexicalist reasoning would go for 'teaching yourself':

- Teaching yourself is doing the same thing to yourself that you do to another when you teach another.
- If A teaches B about *x*, A knows about *x* and imparts/transmits this knowledge to B.
- Therefore, if A teaches himself about *x*, A knows about *x* and imparts this knowledge to himself.

I have assumed a definition of what it is to teach someone something here, which I hope is plausible. If so, then we may note that our use of the lexical formula has landed us with a 'paradox of self-teaching' analogous to the notorious 'paradox of self-deception', for it seems as though the one who teaches himself must, as teacher, know about *x* and at the same time, as student, be ignorant of *x* (and this time it does really seem like these two conditions must hold of the subject simultaneously).

As regards analyzing the notion of being self-taught, the different consequences that would ensue from the adoption of the lexical approach, as opposed to the example-based approach, are clear and striking. Adopting the former, we get the outlandish result that teaching yourself is imparting your own knowledge to yourself. But this answer clearly doesn't tally with the actual use of the expression. That use is simply not constrained by the strictures of any such formula. The cases we refer to with that phrase are not cases in which people impart their own knowledge to themselves (if that were intelligible). They are cases in which people *lack* the relevant knowledge, but acquire it without the benefit or help of a teacher, by for example solitary study, practice, trial and error, and so on. Adopting the example-based approach would give an answer like that to the question of what it is to be

self-taught. Adopting the lexical approach, we would probably end up being skeptics about the possibility of teaching oneself, or we might construct elaborate metaphysical theories to explain how it is possible to impart knowledge to yourself, all of which would be a surreal reflection of the self-deception debate.

It also seems clear that the lexical approach gives us the *wrong* answer for this case. For nobody wants to say that in our actual use of that expression we are *mis*using it, because we are referring to cases that are not cases of imparting knowledge to oneself. Therefore, the application of the lexical formula to the case of self-teaching constitutes a *reductio ad absurdum* of the lexical approach, conceived of as a general approach to the analysis of constructions of the form 'self-*V*'.

The assumption that self-deception has to be closely analogous to interpersonal deception is based on a specious fallacy, one that we might call the 'reflexive fallacy'. This is the fallacy that deceiving yourself has to be doing the same thing to yourself that you do to another when you deceive another. That this is a fallacy can be seen by applying the same reasoning to grammatically similar cases such as teaching yourself, where it yields clearly unacceptable results. At this point the move may be made by the a priori intentionalist to retreat towards a notion of *literal* self-deception. He may claim that although in an extended sense of 'deceive yourself', you need not do the same thing to yourself that you do to another when you deceive another, to *literally* deceive yourself you must (McLaughlin 1996: 45). But it is not clear that this is a valid move at all. After all, when we speak of someone as being self-taught in some area, without implying that he did the very same thing to himself that he does to another when he teaches another, don't we mean to say that he is literally self-taught? 'Literal' is standardly contrasted with 'metaphorical', and we are hardly being metaphorical, or helping ourselves to poetic license with language, in such instances.

This is not to say that self-deception need not bear *any* similarity to interpersonal deception. Surely some similarities must exist between self-deception and other-deception (as well as the other main species of deception: appearance-deception, i.e. that kind of deception in which one can be deceived by appearances) in order to tie them together as species of deception, and to make it intelligible why we would have given them this common title. As Patrick Gardiner has remarked, when we examine instances typically taken to exemplify self-deception, 'we shall (I suspect) find analogies and similarities with cases of deception proper that are sufficient to make the reflexive extension of the concept appear, within limits, reasonably appropriate. But the instances themselves will form a variegated spectrum, and the analogies can in any event never be more than partial ones' (1969-70: 243). And as I will later argue in the final chapter, some theories of self-deception, which hold that self-deception may not involve false or unwarranted belief, do sin against this principle by severing the main common threads between self- and other-deception. However, the similarity clearly needn't be as close at the lexical formula would imply.

### *8. Need we posit intentionality to account for the self-deceiver's responsibility for his/her self-deception?*

Another argument worth mentioning for the logical necessity of an intention to deceive for self-deception, different from those above, concerns the issue of personal *responsibility* for self-deception. It goes like this. If the self-deceived condition is not brought about intentionally, then the subject cannot be held responsible for her own self-deception. But we do deem self-deceivers (at least to some extent) responsible for their self-deception, and rightly so. Therefore self-deception must be intentional (see Steffen 1986: 33-34). This argument is based on the assumption that people can only be held responsible for things that

they do intentionally, and not for things they do unintentionally. This premise is, however, mistaken, as Kent Bach has already pointed out (1981: 368). Consider cases of negligence. Say that you are driving a car, and are not paying sufficient attention to the road because you are conversing or messing about, or not concentrating enough. Then you have an accident and hit Jones. In this case, you may not have intentionally or deliberately hit Jones, but nevertheless, you are responsible for having hit him because you were not paying proper attention to what you were doing. You are guilty of negligence. Of course, your actions are not as reprehensible as if you were to deliberately mow Jones down, but they are blameworthy nonetheless. It may be, then, that self-deception is unintentional, but that self-deceivers are nevertheless guilty of a kind of negligence in their reasoning. They may neglect to guard against and keep in check their tendency to reason in a biased fashion when they are assessing an issue that they have a stake in. They neglect to keep mindful of how their desires and emotions can unduly influence their thinking. This, moreover, explains the fact that the level of censure that self-deception attracts is less severe than that associated with deliberate deception. It explains why our feelings towards the self-deceiver are often a mixture of contempt and pity. This is a normal attitude to take towards those guilty of an offense due to negligence.

## *9. Concluding Remarks.*

In summary, we have examined a number of arguments for the logical necessity of the intentionality condition for self-deception and have found them wanting. Thus we have defended the claim that the conditions given in Mele's analysis are sufficient for self-deception. Considering now the Basic Scenario situations studied in chapter 2, which Mele's analysis captures, in that case we can presume that if the non-intentionalist interpretation of

these cases were the right one, this would not undermine their right to be classified as self-deception. The fact that these cases are ones of self-deception does not hang on whether an intentionalist or non-intentionalist interpretation of the biased behaviours present in these cases is right. But this, of course, is not to give any reason for preferring a non-intentionalist interpretation of these cases over an intentionalist one. This issue will be addressed in the next two chapters where I will criticize the hypothesis that intentions to deceive are operative in the Basic Scenario cases which we have taken to be paradigmatic of self-deception.

# Chapter 4: The Selectivity Problem

## *1. Introduction.*

In chapter 2, two different interpretations were given of the intentions with which the self-deceptive actions are done in Basic Scenario cases, an intentionalist, and a non-intentionalist one, though reasons for preferring either were not given. In chapter 3, it was argued that if the non-intentionalist interpretation of these cases is correct, that would not undermine their right to be classified as cases of self-deception. Now we shall begin looking more at what is to be said for or against either interpretation. In this chapter, we will look at an argument that some intentionalists use to the effect that positing intentions to deceive is indispensable for *explaining* Basic Scenario cases, which is to say, cases where people end up believing that *p* unwarrantedly because of a desire that *p*. Thus we will here again be on the defensive on the behalf of non-intentionalism. In the next chapter I will begin taking a more offensive approach, where I will question whether it is coherent to suppose that the actions done in Basic Scenario cases could be done with an intention to deceive.

The argument in favour of intentionalism I wish to look at here is what Bermúdez calls an 'inference to the best explanation' (2000: 315). The thought here is that 'we cannot understand an important class of psychological phenomena which would normally be labeled self-deception without postulating that the subject is intentionally bringing it about that he come to have a certain belief' (2000: 315). So this is an argument to the effect that the idea of intentionally deceiving oneself is *explanatorily* necessary, rather than logically necessary, for self-deception i.e. the only or the best way to explain how people end up believing unwarrantedly that *p* when they desire that *p* is to suppose that they intentionally bring it about that they believe that *p*. We may call such intentionalists 'explanatory intentionalists'. This argument proceeds by attributing a *selectivity problem* to non-intentionalist attempts to

account for these cases, a problem that, it is argued, can be best overcome by assuming intentionalism.[33]

*2. The Thesis that Intentions to Deceive are Explanatorily Necessary for Self-Deception.*

The selectivity problem charges that non-intentionalists cannot adequately explain the apparently intelligent way in which self-deception occurs. It is alleged that the non-intentionalist theory cannot account for the fact that in every situation in which we desire that *p* we do not form the belief that *p*, especially in those situations when it would be obviously costly or dangerous to have such a false belief (see Talbot 1995. Bermúdez 2000, 1997). One such situation is given by Talbot, who discusses an occasion when he noticed when driving that his brake pedal could be pushed closer to the floor than usual. This produced the anxious fear that his brakes might be failing. Though Talbot anxiously wanted it to be true that his brakes were working, this desire did not lead to him to believe that the brakes were working (1995: 60-61). And thankfully it didn't, since that would be a dangerous belief to have in the circumstances. It would have prevented him from taking whatever preventative or precautionary steps were necessary to avoid an accident. And it seems true, according to Talbot, that people rarely if ever end up with self-deceptive beliefs in similar situations when it would be so obviously costly or dangerous to have that false belief.[34] A typical example of

[33] The sense in which a priori intentionalists 'favour intentionalism' is quite different from the sense in which explanatory intentionalists favour it, and it would be wrong to see these two kinds of theorist as necessarily being allies. As I mentioned before, some a priori intentionalists are skeptics about self-deception, thinking it can't possibly obtain. Their intentionalism is simply a thesis about the *concept* of self-deception. Explanatory intentionalists can't be skeptics about self-deception, since they are using the hypothesis of intentional self-deception *to explain a real life phenomenon*.

[34] This may be open to doubt. Many of the cases often mentioned in the literature are cases where one's self-deceptive belief is quite a dangerous one to have, such as the often mentioned case where one convinces oneself that one is not seriously ill against evidence to the contrary (a misjudgment which could prevent one from taking

self-deception would be the tyrant, or tyrant's minion, who refuses to apologize for his crimes after deceiving himself into thinking that his tyranny was necessary or in the best interest of the nation, or that his corrupt takings were his just deserts, and such false beliefs do not put the believer in any imminent danger. An adequate account of self-deception, then, must explain why self-deception occurs in these cases when it would not be obviously or immediately costly to have a false belief, and not those Talbot-type cases, when it would be obviously costly to have the false belief.

Intentionalists have accused non-intentionalists of leaving it mysterious in their account why the non-intentional biasing processes do not seem to operate in the Talbot-type cases. For how can what Talbot calls a 'non-intentional mechanism' that is 'activated' or 'triggered' by a desire (1995: 59 & 63), differentiate and 'select' so intelligently between occasions when it's costly to have a self-deceptive belief, and occasions when it isn't? These intentionalists themselves insist that they have the answer for what explains this selectivity, an answer which makes use of the thesis that self-deception is the outcome of practical reasoning. The answer goes like this: With the tyrant, he initially realizes that he is a self-serving abomination, an idea incompatible with his self-aggrandised self-concept. He then reasons that on-the-whole he will avoid distress if he makes himself believe that he was only doing what was necessary (this thought may occur in an instant, or may, in Talbot's view, be unconscious). Therefore, it may be prudentially rational for him to, if he can, induce this belief in himself. However, in Talbot's case, it was obvious to him that all things considered, he would *not* have, on-the-whole, avoided most pain by inducing the belief in himself that his pedal was in fine order. It would clearly not have been prudentially rational for him to make himself believe that the pedals are in fine order. Therefore, he didn't deceive himself, (and

---

early action to solve the problem). These cases are fictional however, and are only alleged to be representative of a typical case of self-deception. Though empirical data concerning real life cases of self-deception would have to be gathered to properly settle this matter, we may go along with Talbot's suggestion for now.

probably didn't even think of deceiving himself). In the tyrant's case, however, to use Talbot's words, 'the Expected Utility of interfering with [his] cognitive processes to bias them in favour of *p* exceeds the Expected Utility of not interfering' (1995: 63), and so the tyrant opted for 'interfering with his own cognitive processes'. This, for the intentionalist, is the simple explanation for why self-deception occurs in these kinds of cases, and doesn't occur in some of the other kinds of case. If self-deception is something people do intentionally, then it is up to them to refrain from doing it when it's not to their advantage to do so.[35]

The intentionalist lays down a challenge to the deflationist on the selectivity issue, and throws the ball firmly into the non-intentionalist's side of the court. She questions whether the non-intentionalist has the resources to explain the selectivity of self-deception within a non-intentionalist framework. A number of philosophers have risen to this challenge, and I will detail some of these responses soon. But first it is worth making some comments about some background assumptions that might be informing this critique of non-intentionalism in some cases. For it is possible that this criticism is informed by a misunderstanding of non-intentionalism of the kind that I warned against in chapter 1. This misunderstanding sees the non-intentionlist as advocating the idea that the unwarranted belief is caused or sustained, not as an outcome of any intentional, voluntary actions, but through the operation of non-intelligent, non-intentional, cognitive *mechanisms* which although perhaps purposive[36] are involuntarily triggered by the desire. For if the desire causes the

[35] Talbot seems to wish to construe self-deception as evolutionarily adaptive, in order to explain how it may have survived the process of natural selection (1995: 66). And for it to be adaptive, it must be selective and not blind. And selectivity is best explained by supposing self-deception to be intentional, that is, as being done with the intention of bringing about the most rational outcome.

[36] I mean 'purposive' in the sense in which the mechanism that makes a sunflower turn to the sun is purposive. We may have evolved so that our desires initiated such mechanisms because for some reason, this facilitated the flourishing of our species.

belief by activating or triggering some involuntary mechanism, then one could expect this process to be insensitive to the peculiar circumstances of the self-deceiver, and one could then expect self-deceptive beliefs to occur in Talbot-type cases as much as any other (and Talbot and Bermúdez, the two philosophers who raise the selectivity problem for non-intentionalism, seems to me to be particularly inclined to view non-intentionalism in a non-agentive way). This could be why Talbot thinks that non-intentionalism would predict that self-deception would occur in the brake-malfunction case. And in fairness to Talbot, there are, as we have seen, indeed *some* non-intentionalists (who I have called non-agentive deflationists), who suggest precisely this picture (e.g. Johnston 1988), though this is not the kind of non-intentionalism I have been recommending. For as I have argued, the non-intentionalist *is* at liberty to admit voluntary, intentional actions into his account, and to regard them as the primary drivers of self-deception, so long as the associated intentions are not intentions to deceive.

Be that as it may, a number of non-intentionalists have taken up the challenge posed by the selectivity argument, notably, Barnes and Mele. Barnes says that we have a universal propensity to be partial towards anxiety-reducing beliefs, but that this propensity may be overridden by other propensities in certain circumstances. Responding to Talbot's example she says:

> It seems both possible and plausible to think that other universal dispositions exist which, in many circumstances, override the universal disposition to be biased in favour of anxiety-reducing beliefs. People are, for example, universally disposed in the face of danger to take action to protect themselves. If I have an anxious desire that my car brakes be in good working order, my desire to protect myself from being injured in a car accident typically overrides my partiality for beliefs that reduce my

> anxiety. I bring the car in to be checked and serviced rather than deceive myself into believing that my brakes are in good working order (1997: 81).

So for Barnes, although the tendency for these biasing processes to operate is always there, in certain circumstances the tendency is overridden by other stronger dispositions: '…everyone in certain situations of anxious desire is disposed to be partial to anxiety-reducing beliefs. Self-deceivers are those of us in whom this disposition to be partial is not overridden, or resisted, or otherwise prevented from operating' (Barnes 1997: 82).

Mele gives a similar but more developed kind of answer to that of Barnes, exploiting his FTL model which we looked at in chapter 2. To remind ourselves, the FTL model uses the notion of a 'confidence threshold', which refers to the quantity and quality of evidence required to convince one that something is true. This model proposes that confidence thresholds may vary from case to case depending on what the expected costs are of *falsely* believing the relevant proposition. The less costs expected to be associated with falsely believing that *p*, the lower one's confidence threshold regarding it (i.e. it will take evidence of relatively lesser quality or quantity to make one believe it), and the more costs expected to be associated with falsely believing that *p*, the higher one's confidence threshold (i.e. it will take evidence of a relatively higher quality or quantity to convince one of it).

Mele claims that this model has the power to explain the phenomena Talbot has in mind. He illustrates this with the fictional case of a CIA agent, Gordon, who is accused of treason. Two interested parties, who have approximately the same information about Gordon, form beliefs about whether the accusation is true or not: his parents, and his work colleagues. Both parties want it to be untrue that he committed treason, and to the same degree. His colleagues want him to be innocent because it would mean they are at less personal risk, and they are quite fond of Gordon, and the reasons why his parents want him to be innocent

should be obvious. However, while his parents end up believing, unwarrantedly, that Gordon is innocent, his colleagues end up believing truly that he is guilty.

Mele explains this difference using the FTL model, in terms of the different costs associated with believing falsely that Gordon committed treason. For his work colleagues, the costs of falsely believing that he is innocent are very high, in that their safety may be at stake if he really were guilty (if it is true that he is a double-agent, for instance, then he may be passing on details about them to the enemy). This increases their confidence threshold, such that they will require relatively good evidence before they accept the hypothesis that Gordon is innocent. With Gordon's parents, on the other hand, there may be no analogous costs associated with believing falsely that Gordon is innocent, which sets their confidence thresholds lower than that of his work colleagues. It then will take less evidence, or evidence of a lesser quality, before they accept the hypothesis that he is innocent. Note that this explanation is consistent with that of Barnes. To use Barnes' terminology, the general disposition that these work colleagues have towards anxiety reducing beliefs is overridden in this case by their fear of believing falsely that Gordon is innocent, a fear that is lacking in Gordon's parents, and that is why they don't end up self-deceived, while the parents do. Similarly, in Talbot's case we can see that what prevented him from forming the welcome belief that his brakes were working fine was his appreciation of the costs of falsely believing that proposition if it were false, and his desire not to miss the opportunity to take remedial action. If Talbot was disposed to be biased in this situation, this attitude-complex would have made him disposed to be biased in favour of forming the unwelcome belief, rather than the welcome one.

I hope this gives some indication that a coherent and plausible response is possible to the selectivity problem within a non-intentionalist framework. I believe that the empirical evidence we looked at in chapter 2 shows that people have a propensity towards reasoning in

a biased way when they encounter evidence that is very distressing. They have a propensity towards seeking evidence favourable to their desired conclusion, while failing to make proportionate efforts to seek unwelcome evidence. Furthermore, they are also disposed to jump to the welcome conclusion sooner than would an impartial judge of the evidence if they have a low confidence threshold for believing the welcome proposition. These two factors may conspire together to cause one to have a biased, self-deceptive belief. But it is not inevitable that these things happen, and we can work to keep this tendency in check. When there may be risks associated with having such a comforting belief, we will be particularly motivated to exercise caution and keep this tendency in check, and this is why self-deception may happen in some types of situations but not others.

### *3. Concluding Remarks.*

Intentionalists such as Talbot and Bermúdez claim that non-intentionalists have no plausible way of explaining the 'selectivity' of self-deception. However, non-intentionalists like Barnes and Mele have offered seemingly plausible ways of doing this, while staying within a non-intentionalist framework. These deflationary responses seem to throw the ball back into the intentionalists side of the court on this issue. However, whether Barnes' and Mele's treatment of this issue is ultimately satisfactory or not, I believe that some satisfactory non-intentionalist response to the selectivity problem must be available, since as we shall see next, the supposition that intentions to deceive are operative in Basic Scenario cases is scarcely coherent, and hence by a process of elimination, a non-intentionalist characterization of the intentions involved in Basic Scenario cases must be correct.

# Chapter 5: Intentional Action and Knowledge of What you are Doing

## *1. Introduction.*

After having been for quite some time on the defensive on behalf of non-intentionalism, I will now take up a more offensive approach. For the question has not yet been asked: what exactly is it that the non-intentionalist finds objectionable with intentionalism, such that he/she is motivated to seek an alternative account? The answer, in short, is that the idea of intentionally deceiving oneself has been usually deemed worth avoiding because it has certain paradoxical consequences, and because it is regarded as something it would be impossible to realize (in ordinary circumstances), which would preclude it from being of any explanatory use with respect to the kinds of cases we are examining. Next we will examine why exactly one might think of intentional self-deception as paradoxical and impossible to realize in ordinary circumstances. The reason, it will emerge, is based on the thesis that doing something intentionally entails doing it knowingly. I call this the *knowledge condition* on intentional action.

Regarding this knowledge condition, we should ask, (1) is this a condition that we should accept?, and (2) if this should be accepted, will it rule out an intentionalist interpretation of common Basic Scenario cases? My conclusion will be in the affirmative on both counts. I will argue that doing something intentionally entails doing it knowingly, and that if we then accept the knowledge condition as a general constraint that an intentionalist theory of self-deception must respect, then *for a number of* (though not all) prospective strategies of intentional self-deception, this constraint would mean that one could not possibly deceive oneself intentionally by such means. And in particular, for the kinds of actions seen to be operative in the Basic Scenario cases, this constraint would prove fatal for the prospects of an intentionalist interpretation of those actions. Therefore, by a process of

elimination, this will amount to a recommendation for the non-intentionalist interpretation of these intentional actions. This I will attempt to show in this chapter.

Let me just state now though that this will not be the end of the matter. For as I will discuss in chapter 6, there is one particular prospective strategy of intentional self-deception which may be immune to these objections. That is, there may be one strategy of intentional self-deception which could be carried out knowingly and yet successfully. However, this particular strategy seems to be not particularly relevant to the question of how to interpret common Basic Scenario cases, since psychologists have found no evidence that it is operative in such cases. Nevertheless, I will show that it is open to the intentionalist to argue that it's possible that this strategy is undertaken in cases which are similar to garden-variety Basic Scenario cases, but which have escaped the scrutiny of experimental psychologists for various reasons. Thus the intentionalist can try to carve out an area for herself for which her intentionalist theory has application. The strategy I have in mind here I call the *attentional strateg*y, which involves shifting attention between belief-relevant considerations. And as I will argue, what makes it crucially different from other potential strategies of self-deception, such that it might overcome the knowledge condition objection, it that it is a *knowledge undermining* strategy. That is, it is specifically geared towards suppressing any knowledge one may have which is standing in the way of one having the desired belief. Discussion of it will be reserved for the following chapter, where that strategy will be criticized on empirical grounds, and hence on grounds that are different from that of demonstrating any paradoxes associated with it.

*2. Intentional Action and Awareness of What One is Doing.*

Deflationist skepticism over the viability of intentionalist accounts has its source in the following point mentioned by Mele:

> It is often held that doing something intentionally entails doing it knowingly. If that is so, and if *deceiving* is by definition an intentional activity, then one who deceives oneself does so *knowingly*. But knowingly deceiving oneself into believing that *p* would require knowing that what one is getting oneself to believe is false. How can that knowledge fail to undermine the very project of deceiving oneself? It is hard to imagine how one person can deceive another into believing that *p* if the latter person knows exactly what the former is up to. And it is difficult to see how the trick can be any easier when the intending deceiver and the intended victim are the same person (1997a: 92).

Mele dubs this problem associated with the idea of intentionally deceiving oneself, the *dynamic paradox*. Fundamental to this paradox is the presupposition that if one *V*ed intentionally, one necessarily must have *known* or *been aware* that one was doing *that*, namely, *V*ing.[37] In other words, the description of the action under which it was done intentionally is a description that you must have known was applicable to the action when you did it. This seems to suggest that 'unconsciously yet intentionally *V*ing' is contradictory.

Philosophers have frequently thought that there's a conceptual connection between the notion of intentional action and knowledge/awareness of what you're doing (see

[37] Actually, whether Mele is really committed to this is unclear, since he states elsewhere that 'hidden intentions' are possible (1997: 100). However, Mele does think that there is something paradoxical about the idea of intentionally deceiving yourself, and it's hard to see what else could be generating this paradox if not this presupposition. I am in agreement with Bermúdez on the issue of this alleged entailment, who says that 'nothing less than this will generate the dynamic paradox' (1997: 108).

Anscombe 1957: 11 & 87. Bratman 1984: 387. Donnellan 1963: 406. Gorr and Horgan 1982. Gustafson 1975: 89. Hamlyn 1971: 46. Hampshire 1970: 145. Moran 2001: 125. Mele and Moser 1994: 41-42. Miller 1980: 334). Some lexicographers apparently assume so too. For instance, the *American Heritage Dictionary*, 4th edition, defines 'deliberate', which is often used interchangeably with 'intentional' in the self-deception literature, as 'done with or marked by full consciousness of the nature and effects; intentional'. I assume the relevant considerations that might support such claims would be as follows. Take any occasion when you do something without knowing or being aware that you are doing it. Say, for instance, that you are making soup, and you unwittingly add some sugar (you think that it's salt). Or imagine that you travel to an exotic country with very different customs and you meet one of the locals. You do something that would not attract any notice in your country but which is taken to be insulting in this culture, causing offence to the local, though you weren't aware that this action is considered offensive. Now it seems clear that in these cases we say that we added the sugar or offended the local neither intentionally nor deliberately, and in defending this claim, we naturally advert to the fact that we weren't aware that that was what we were doing, which we take to be sufficient to show that we couldn't have done these things intentionally. Saying that would have presupposed awareness that the substance was sugar, or that the action could be construed as an insult. A philosopher could then argue that by parity with such cases, if someone deceived herself intentionally, she must have known that she was doing that.

This is not to imply that if an agent does something unintentionally, there are not some descriptions of the action under which it was intentional and under which the agent was aware of herself doing it. Consider the following descriptions of an action that I may perform at an auction: 1) I raise my finger, 2) I bid for the painting, 3) I bid for the forgery. Let's say, however, that I didn't know that the painting was a forgery, though I knew that I was raising

my finger and bidding for a painting. In that case, the claim is that I did not, and indeed *could not* have *intentionally* bid for the forgery, though I knew that I was bidding for a painting and bid for it intentionally.

It is also worth noting that, at least if the surveys of Malle and Knobe (1997: 112-114) are anything to go by, the judgments of philosophers in these cases are in accord with the 'intuitions', or linguistic usage, of ordinary people too. From the use of questionnaires describing certain actions given to 225 undergraduate students, Malle and Knobe found that they were generally unwilling to attribute intentionality to the actions if the actors were described as being unaware that they were doing them.[38] They conclude that to the theorist 'who believes in unconsciously performed intentional actions (dispelling the awareness component), the results must be surprising, if not damaging' (1997: 114).

*3. Objections.*

As one might expect, objections to this thesis have been raised, and purported counterexamples brought forward. This thesis, which is crucial to the critique of intentionalism that follows, will now be defended against them. Some of these objections will force qualifications in our position, but I will argue that they do not amount to a refutation of the knowledge condition thesis.

Objection 1) Firstly, consider a case where the success of our intentional enterprise is not a sure thing (Davidson 2001/1978: 91-92. Ross 1982: 264). Imagine NASA launching a rocket programmed to destroy an asteroid hurtling towards Earth. Imagine also that they have serious doubts that the mission will succeed though the exigency demands they try, and

[38] They also found *skill* to be an important component of the concept of intention, in that subject's need to have the skill to pull off an action in order to be able to do that action intentionally, though this point is beyond our concern here.

imagine that they do succeed. Because of these doubts, the argument goes, we may not be able to describe them as having *known* that they would destroy the asteroid (if such uncertainty is incompatible with knowledge), yet it seems correct to say that they intentionally destroyed the asteroid. This is an important exception, yet we would be overreacting to it if we were to deny, on its basis, any necessary link between intentionally action and awareness of what you're doing, since surely in all cases like this, the agents would have to be aware that they were at least *trying* to do what they succeeded in doing. It's hard to imagine how NASA, for instance, could have intentionally destroyed the asteroid without having known that that's what they were *trying* to do. These kinds of case, then, certainly give us cause to modify our statement of the knowledge condition:

> Knowledge condition: If one *V*s intentionally, one must know or be aware that one is *V*ing, or at least that one is trying to *V* (in cases where one doesn't know if one's attempt at *V*ing will be successful).

But this qualification is surely no admission that there can be unconscious intentional action in the sense that would be congenial to intentionalism.

Objection 2): Ross considers a purposive action 'done by rote' and 'mechanically', like shifting into third gear during the course of driving, and says, '[m]y intuition is that such an act is intentional even though one does not at the time [know]…that one has performed it' (1982: 263) and others too have expressed similar views (e.g. Martin 1997. Fingarette 1998). I would imagine that the reason why Ross assumes that in such cases the subject didn't know at the time that he was doing this action is because he's imagining that the subject was *not aware* of himself doing the action, his attention being engrossed in other things. However, whether people really are unqualifiedly unaware of what they are doing in such cases is

debatable. For awareness, it seems, need not be an all or nothing affair, but may come in degrees, and it seems that such cases would not untypically involve partial or peripheral awareness of what one is doing (that is, it seems as wrong to say, unreservedly, that subjects are *completely unaware* and *oblivious* to what they are doing in such cases as it is to say they are fully aware of what they are doing). Granted, in light of such peripheral awareness we would be as reluctant to say, unqualifiedly, that the subject knew she was shifting into third gear as we would be to say that she didn't know this, but this reluctance is proportional with, and explains, our hesitation on the question of whether to classify this as an intentional or unintentional action. It is an exemplar of neither. In other words, the fact that there is typically partial awareness of what one is doing in such cases explains the (reserved) inclination, where it may exist, to want to say the action was intentional, so this case does not clearly demonstrate a disconnection between intentional action and knowledge of what you are doing.

Now it might be thought that these kinds of actions would be a good model for understanding the intentionally deceptive act supposedly involved in self-deception. That is, perhaps the self-deceiver's actions are intentional in an attenuated sense, and they are hazily aware of what they are doing, and because of this hazy awareness, they are able to pull it off successfully. But this does not seem plausible because it is difficult to understand how the self-deceiver would be hazily aware of her self-deceptive actions. The reason for this is that the explanation for why we have only peripheral awareness of what we are doing in the uncontroversial cases doesn't seem applicable in this case. In the driving case, for instance, we typically fail to be fully aware of our gear shifting because we are so used to doing it that we need not think about it, and because our minds are attracted towards things that are of more interest or significance for us, such as thoughts of this and that, or a conversation with a passenger. However, our deliberate act of self-deception is not like this. As the typical cases

are described, in the issue surrounding our deceiving ourselves is *the* significant issue for us in that moment; it is laden with affective significance (this point is made by Poellner (2004: 57-60) who uses it to criticize Sartre's view that an intentional act of self-deception can be carried out 'pre-reflectively'). And furthermore, we can't say that, like shifting gears, it is something we could do unconsciously because we are 'so used to doing it we need not think about it'. For then we could ask about that time when you *started* to learn how to deceive yourself: 'how did you manage to do it with this casual, semi-awareness *then*?', and now the answer isn't available that you were so used to doing it you didn't have to think about it. There is no problem, of course, with the idea of the driver having to pay attention to his driving before he learns how to do it automatically, but there is with the idea of the intentional self-deceiver having to pay attention to his deceptive activity before learning how to do it automatically.

Perhaps this point is unfair because, as the intentionalist taking this line may argue, self-deception is not a special skill like driving which one needs to pay close attention to initially in order to become adept at it. This may be so, but it seems that if the intentionalist wants to insist that the self-deceiver is not fully aware of his actions under the description of 'deceiving myself', he still owes us an explanation for why this is. Such an explanation is available in the shifting-into-third-gear case, but one would have to make good on the promise for an equivalent kind of explanation to convince us of this partial awareness thesis, especially considering how implausible the idea seems in light of the affective significance associated with the self-deceptive project for the self-deceiver.

Objection 3) Another philosopher who challenges the idea of the knowledge condition is J.L. Bermúdez, who says that 'it seems false that one cannot do something intentionally without doing it knowingly' (2000: 314). Bermúdez is not here denying that an action would

have to be *initiated* by the person knowing the intention with which he is doing it. Rather he says that:

> …one can lose touch with an intention while one is in the process of implementing it, particularly when that implementation is a long drawn out process. The fact that an action is precipitated by a conscious intention does not entail that while carrying out the action one remains constantly conscious of the intention that gave rise to it. By the same token, the fact that one is not conscious of the intention while carrying out the action does not undermine the action's status as intentional (2000: 314).

This criticism foists on the advocate of the knowledge condition an unnecessarily strong claim, however. It is not a requirement of this thesis that the person 'remains constantly conscious' of the intention with which he is doing what he's doing, since it is not implied by the concept of knowledge that he needs to be. A wage-earner may go off to work with the intention of earning a wage, and sure enough, he may not be thinking about that goal then or throughout the day, yet he still knows his intention in doing what he is doing in the sense that if you were to ask him what his intentions are in doing it, he would be able to say that it is to earn a wage. For him to know his intention in what he is doing he needn't be constantly thinking of it; he just should not have forgotten it.

Let it be clear that this thesis need not imply that there cannot be unconscious intentions. It is a thesis specifically about intentional *actions*. One can have an intention and not act on it, and we need not insist here that such intentions cannot be unconscious. If, for example, an intention which I have forgotten (especially due to repression, if there is such a thing) and which is hence inaccessible to my consciousness is counted as an unconscious intention, we need not put up an argument. However, a forgotten/repressed intention is not

one that we act with, and the claim here is only that one cannot *do* something *with* a certain intention without knowing what one's intention is. One cannot act on an intention while it remains inaccessible to consciousness. However, even for those who are inclined to accept that it makes sense to speak of doing something intentionally but unknowingly, an important objection awaits that has been voiced by Lazar with respect to the case of self-deception in particular (1999: 277-279). I will explain this to the best of my understanding.

### *4. Lazar's Criticism of Unconscious Intentional Action in Self-deception.*

Before we leave the topic of defending the knowledge condition, it is perhaps worth mentioning a criticism of the idea that unconscious intentions are operative in self-deception made by Ariela Lazar. It is an argument that is, if not valid, then at least interesting and deserving of a response by the intentionalist advocating the role of unconscious intentions in self-deception. This criticism begins by taking the logical possibility of unconscious intentional action for granted and then working from there towards a *reduction ad absurdum* of sorts.

Consider first a relatively well worked-out account of how unconscious intentional action features in self-deception, that of W.J. Talbot. According to Talbot, self-deception is intentional, and is accomplished by the subject's intentionally 'interfering with' or 'biasing her cognitive processes'. For Talbot, this can mean things like making it such that one remembers selectively (in that one remembers things congenial to the favoured belief and doesn't remember things uncongenial to it). So it may involve selectively making or keeping memories of unwelcome evidence in an inaccessible or less accessible state. For Talbot, however, if the subject was aware of the intention with which she is acting, then this would undermine the belief, because one cannot hold a belief that *p*, while knowing that this belief

was formed from biased reasoning, or through an intention to deceive. Therefore, the intention, he supposes, must itself be unconscious. We may take this to mean that the subject is unaware of the intentional action itself being done, or that she is aware of the action, but under some other description.

But intentions being unconscious would not be the norm for intentions, and so if the intention is unconscious then there must be some explanation for why it is. The question of how the intention ended up being unconscious, however, is a question that Talbot (as well as the other thinkers who implicate unconscious intentions in self-deception) neglects to address. Lazar mentions two possible kinds of explanation for why something can be unconscious, a 'non-thematic' one, or a 'thematic' one.

As I understand it, the distinction is as follows. Something (a perception, attitude, or an action) can be unconscious either because the subject made it so for a particular reason usually relating to the 'content' of the thing itself and how it relates to the subject's goals, or it can be unconscious for some reason other than her making it so. Examples of the latter are widespread. In cases of subliminal perception, or visual masking, a stimulus might be perceived but might not reach consciousness. Here, the explanation for why the perception was unconscious has nothing to do with the subject making these things unconscious, but has to do with general facts about the viewing conditions and human 'information processing' capacity. Or (to take a stock example) someone might not be aware (or at least be only partially aware) of her actions of driving the car or of the road in front of her, because she is so engrossed in a conversation with her passenger. Here, she has not made herself unconscious of these things, rather, she is unconscious of them because the resources of attention are limited and are needed elsewhere, and the actions can be carried on habitually. In non-thematic explanations, the nature or 'content' of the thing does not figure in the explanation of why it is unconscious. For instance, so long as a stimulus is shown to a subject

at a rapid enough speed it will be unconscious, no matter what the stimulus happens to be. In such cases the thing being unconscious is 'explained by general facts and principles pertaining to information processing' (1999: 278).

As Lazar then suggests, it would be very implausible to claim that there is a non-thematic explanation for why the intention to deceive is unconscious in the case of self-deception, pertaining to 'general facts and principles of information processing'. Self-deception for the intentionalist is essentially thematic. The intention to deceive is supposed to be unconscious because of the particular intention that it is and how its being conscious would interfere with the subject's strategy. It is not unconscious for some reason not related to its significance for the self-deceiver's goals. But in that case, it must be the subject herself who is responsible for it being unconscious, since it is only her who understands the significance of the intention with respect to her own goals.

However, if the explanation for why the intention is unconscious is that the subject intentionally made it unconscious, then we have another intention to make the first intention unconscious. We may then ask of this second-order intention whether the subject is conscious of it or not. If the subject is conscious of it, then this would threaten to unravel the self-deceptive project no less than if the first-order intention were conscious. But if the second-order intention is itself unconscious, then we are going to need an explanation for why it is unconscious, which if Lazar's arguments are right, can't be non-thematic. Therefore we will require a third-order intentional act to make this second-order intention unconscious. And so on *ad infinitum*. And we need no argument to say that such an infinite chain of intentional acts would not be an explanatory option.

At this stage I will leave the defense of the knowledge condition. Certainly more objections to the knowledge condition may come from unseen directions.[39] But from here on in I will consider the knowledge condition thesis as one deserving our endorsement, and as a constraint that an intentionalist theory of self-deception should respect. As we saw earlier, Mele regards the view that intentional action entails knowledge of what you are doing as amounting to a serious difficulty for the prospect of intentional self-deception, rendering the idea paradoxical. Let's have a look now at how exactly this knowledge would preclude success in any such project with respect to a number of potential strategies of intentional self-deceptive often mentioned in the literature. As we will see, it would preclude success for quite a number of prospective strategies, though there is one popular strategy which would not be in principle impossible under this constraint.

## *5. Strategies of Intentional Self-Deception and the Knowledge Condition.*

It is widely held that one cannot simply make oneself believe something at will, in the sense in which this might be a 'basic action' one could perform. This means that we cannot make ourselves believe something without doing something else by means of which we make ourselves believe it, in the way that I can, say, clench my fist at will, without having to do anything else to do it (the impossibility of making yourself believe something at will is often thought not to be a contingent impossibility, in the way that it's impossible for many of us to twitch our ears at will (though some can do this), but rather, a logical impossibility; the idea is thought to make no sense in light of what beliefs are (e.g. see Williams 1970/1993)). So if one is to intentionally deceive oneself into believing something one knows to be unwarranted, then one would have to do it *by doing something else*, or by executing some *strategy*. This is

[39] One might mention psychoanalytic cases as potential counterexamples here, though I feel they are mired in enough controversy to not warrant discussion here.

no different from our position relative to another with respect to deceiving them: we cannot make someone else believe something at will, but we must employ some strategy there too, such as telling him/her lies. Assuming that what we have said about the connection between intentional action and knowledge is correct, the challenge for the intentionalist, then, is to show us that there is a *viable* or *feasible* strategy by which the self-deceiver could deceive himself, granting the knowledge condition as a theoretical constraint.

Let it be noted, however, that the question of whether there is a viable strategy of intentional self-deception should not be construed so generally so as to include the use of what we might call 'special means'. For philosophers who are most critical of intentionalism can be often found admitting that it would be possible to deceive oneself in the relevant sense through the use of special means, such as brain-altering technology. These possibilities, however, are usually not thought to be pertinent. When Annette Barnes, for instance, inquires into the possibility of intentional self-deception, she says 'I am concerned with what must be realistically done [to intentionally deceive oneself]. Taking a pill, hypnosis, etc. are not realistic options here' (1997: 27). Presumably, the idea is that these possibilities are not pertinent, because they would not vindicate the intentionalist's claim that intentional self-deception is what is going on in ordinary or *garden-variety* cases of self-deception (i.e. Basic Scenario cases), where such special means are not available, and where people only have the ordinary resources of their own minds to rely upon. No one may doubt that one could deceive oneself intentionally if one had available such special means to do so. But in ordinary cases where people form welcome beliefs against good evidence to the contrary, people don't have special means available, and what is questioned is whether in *such* circumstances, the hypothesis that people are intentionally deceiving themselves is an accurate diagnosis of the situation.

Similarly, some of the staunchest critics of intentionalism are frequently found admitting and illustrating the possibility of intentional self-deception with a case that we may call *The Diary Case*, which does not involve the use of any 'special means' as such. In this case, a woman has an appointment in a few weeks that she'd rather not attend. She then hatches a plan to deceive herself into thinking it's on a date that it's not on, so that she'll miss it. She deliberately marks the meeting down in her diary as being on a date she knows it's not on. Because she knows that she has a bad memory, she knows that when she looks at her diary in a few weeks she will have forgotten all about her deceptive plan by then, and will accept what the diary says as accurate (see McLaughlin 1988: 31-32, Mele 1997a: 99. Lazar 1999: 271. Bok 1980: 928). A similar and fascinating example can be seen in the finale of the movie *Memento* (2000, directed by Christopher Nolan), where (beware of spoilers!) the main character intentionally deceives himself by writing a message for himself that he knows to be false, exploiting the fact that he suffers from anterograde amnesia, a condition where one is unable to form new long-term memories. In these cases, the subject's knowledge of what he/she was doing would not be an obstacle to the success of the project, because over time that knowledge may be forgotten, leaving the way clear for the strategy to succeed. As I've said, the intelligibility of such cases would be admitted by all parties to the debate. But these kinds of case are far removed from the idea we have of ordinary cases of self-deception, and the strategies are far removed from the kinds of strategies that have traditionally been implicated in these ordinary cases, such as rationalization, selective attention, and selective evidence gathering. They would, moreover, only be possible for people with very or pathologically bad memories. So we can take the important question to be one of whether it is possible to intentionally deceive oneself using those strategies that have been *traditionally implicated* in self-deception, such as rationalization, etc., strategies of the sort that one might presume to be available to ordinary people, in ordinary circumstances.

Before we take a look at some of those strategies, allow me to illustrate the form of the non-intentionalist's argument against the possibility of intentional self-deception (in ordinary circumstances), with a very simple if unrealistic case. Consider what is perhaps the most common strategy that may be used to deceive another: lying. Consider then if one could deceive *oneself* by turning this strategy upon oneself.

Consider, then, lying to yourself, as this expression might be used literally (and not as a synonym for self-deception, as it is sometimes used). If you lie to yourself with the intention of deceiving yourself, then you lie to yourself intentionally. If you lie to yourself intentionally then you do it knowingly. But if you knowingly utter a lie to yourself, then you know it is a lie, i.e. an untruth. And for you to know that would presumably render you immune to it; you could not be taken in by that untruth. Such an action would be a futile performance.

This illustrates the form that the argument against the possibility of intentionally deceiving oneself takes. But let us now consider some strategies that actually have been implicated in self-deception by philosophers, and ask whether one could succeed in deceiving oneself by intentionally employing these strategies. The two examples I wish to consider are *rationalization* and *selective evidence gathering*.

Consider rationalization first. On Kent Bach's definition, which seems typical, rationalization is 'any case of a person's explaining away what he would normally regard as adequate evidence for a certain proposition' (1981: 358). Here, 'explaining away' is, I gather, meant as a pejorative expression. It does not mean refuting that evidence, or pointing out weaknesses with it, but it means something more like coming up with specious arguments to undermine the evidence, arguments that perhaps have some *prima facie* plausibility, but which are ultimately bogus or flimsy. Assuming this, if Jones rationalizes intentionally, and if rationalizing means adducing specious arguments to undermine evidence, then Jones

intentionally adduces specious arguments to undermine the evidence. And if doing that intentionally implies doing it knowingly, then he *understands* himself to be adducing *specious* arguments. But knowing that, it is difficult to see how he could be convinced by those arguments. Saying that he knows the arguments are specious just seems incompatible with saying that he could find them convincing. Though someone may intentionally rationalize to fool another, the idea of intentionally rationalizing to fool *yourself*–given our assumptions–seems incoherent. One would expect rationalizers not to *take themselves* to be rationalizing, and so not to be intentionally rationalizing.

Consider also selective evidence gathering, a phenomenon that has also been implicated in self-deception. Selective evidence gathering could be where *S* ends up with the unwarranted belief that *p* because she searches for evidence supportive of *p* while avoiding/neglecting to search for evidence supportive of not-*p*. In this way one's evidence search is biased. One ends up collecting a body of evidence which is one-sided. Now if one was to intentionally search for evidence in a biased way with the intention of making oneself acquire an unwarranted belief, then one would have to do this knowingly. However, if one were to know that the body of evidence one has collected is one-sided, then it is again difficult to see how one would be convinced by that body of evidence. Presumably this knowledge would prevent or at least very much limit the extent to which one would be taken in by it. So it seems doubtful that self-deceivers could intentionally or deliberately gather evidence selectively.

### *6. Relevance of the Above Inquiry into the Interpretation of Basic Scenario Cases.*

There is now something quite important for us to take note of. These two behaviours, rationalizing and selective evidence gathering, are of particular significance, since these kinds

of behaviours were seen to be operative in the Basic Scenario situations. In the experiments we looked at in chapter 2, the evidence indicated that rationalization was probably occurring, insofar as it was found that the process leading to the subjects acquiring the biased belief partly involved them being hyper- or unreasonably critical towards the unwelcome evidence. A more specific example of rationalization was also found in the My Lai Massacre case mentioned in the beginning of chapter 1, where subjects unreasonably refused to believe that American soldiers had committed atrocities in Vietnam. They appeared to be rationalizing, insofar as they explained away the evidence (stories and photographs) with such far-from-conclusive hypotheses as 'The story was planted by Viet Cong sympathizers and people inside the country who are trying to get us out of Vietnam', hypotheses which they gave unwarranted credence to. These experiments also found evidence of bias and selectivity in subjects' evidence gathering procedures, broadly construed (where searching in one's mind for reasons counts as evidence gathering).

Now remember that one of the important questions we set ourselves was whether the actions which produce the biased belief in Basic Scenario cases should be interpreted in intentionalist or non-intentionalist terms. And it seems now that insofar as these actions include things like rationalizing and gathering evidence selectively, they *can't* be understood in intentionalist terms, that is, the subject can't be understood as doing these things intentionally for the purpose of deceiving himself. For instance, to describe one of the subjects as *rationalizing* in the My Lai Massacre case in hypothesizing that the story of American atrocities was 'planted by Viet Cong sympathizers and people inside the country who are trying to get us out of Vietnam', is to pass judgment on the evidential worth of this explanation, affirming it to be flimsy, and not deserving much credence. To then say that this person was *intentionally* rationalizing in giving this explanation is to imply that she adduced this hypothesis *believing* it to be flimsy, and this is not a plausible view of the situation at all.

Therefore, the non-intentionalist characterization of this paradigmatic case of self-deception wins by default, given that the alternative, intentionalist characterization of it is scarcely coherent.

The general problem for the intentionalist here is that doing these epistemically improper activities intentionally brings with it knowledge that you are doing or that you did something improper, and this knowledge would prevent you from being taken in by these activities. As Mele puts it (1997a: 92), if I'm aware that you are reasoning in a biased way about some issue, then I will not be disposed to warm to whatever conclusion you reach from that reasoning. And why should things be any different if I believed that my own reasoning was biased? How could I have any faith in any conclusion known to have been reached through that? Unless one was to do something to suppress this knowledge (an idea we will discuss in the next chapter), it would seem to be a serious impediment towards the success of the project. This is why it does not seem like we are onto a winner with the intentionalist's interpretation of the behaviours operative in Basic Scenario situation.

Let us note the nature of the impossibility here. It seems that when we say that it would be impossible to intentionally deceive yourself by intentionally rationalizing or intentionally seeking evidence in a biased way, we are trying to say that the very idea of doing these things intentionally and successfully deceiving yourself is paradoxical or incoherent. The idea that one intentionally thinks up of a specious reason to explain away unwelcome evidence, which one thereby *knows* to be specious, whereupon one finds that reason convincing and thus dismisses the unwelcome evidence, does not seem coherent, for one's finding that reason convincing would be a criterion for one's not knowing that it is specious. You can't find a reason specious and convincing at the same time.

*7. The Homuncular Approach to Solving the Dynamic Paradox.*

There is a kind of response to the problems posed by the dynamic paradox for intentionalism which, though perhaps considered somewhat outdated now, commanded quite a bit of attention in the literature at one stage, and ought to be discussed for the sake of completeness if nothing else. This is the so-called homuncular response to the dynamic paradox. This theory does not propose a certain alternative strategy, to be put alongside the others of rationalization, selective evidence gathering, the attentional strategy etc., which we are examining, an alternative which may not be susceptible to the problems encountered. Rather it recommends radically re-conceptualizing the mind in such a way as to make logical space for the possibility of intentional self-deception, granting the assumption of the knowledge condition. We will examine this response in this section.

The best proponent of the homuncular response to the dynamic paradox is David Pears. Though Pears himself might have objected to his theory being labeled 'homuncular' (in that it conjures up the idea of a little person in the head, an idea he would reject), if we define a homuncular theory as one in which the intention to deceive is ascribed, not to the person, but to a part of the person or something inside the person, then a convincing case could be made for taking Pears' theory to be homuncular.

Again, the problem here is how is one to succeed in intentionally deceiving herself if she must do that knowingly? To answer this, Pears maintains that we should look on the self-deceiver as being composed of two 'systems': a 'sub-system' and a 'main system'. The main system is identical to or closely associated with the person's consciousness (1984: 101). The sub-system, on the other hand, is 'a separate centre of agency within the whole person' (1984: 87) which, though within the person, is 'organized like a person' itself. This sub-system has, according to Pears, a sort of 'altruistic' (1984: 91) concern for the main system,

apparently wanting it not to suffer from the knowledge of the unwelcome truth. Accordingly, it 'wants the main system to form the [contrary] irrational belief and it is aware that it will not form it, if the cautionary belief [i.e., the belief that the irrational belief is irrational] is allowed to intervene' (1984: 87). The sub-system then intervenes, 'suppress[ing]' this cautionary belief (1984: 101), and 'generat[ing] [the irrational belief] directly in the main system' (1991: 204). Pears seriously entertains the idea that the sub-system is itself a conscious entity, though he comes to no definite conclusion on this point (1984, pp.98-100). For one thing, it has to 'compare the outcome it is producing with the outcome it aimed for and act or cease to act accordingly' (McLaughlin 1988: 82).

The sub-system manipulates the main system intentionally, and so we may suppose, knowingly. But this knowledge may not jeopardize the existence of the unwarranted belief, simply because this knowledge is the sub-system's knowledge, while the belief is the main system's, and as Pears supposes, the main system is unaware of the sub-system's activities and designs, and indeed of its very existence (1991: 404). As Pears explains to us, it 'is their separateness that then seems to explain why a belief that is in one of them does not produce its normal effect in the other, in much the same way that the separateness of two different people would explain why a cautionary belief in one of them would not automatically stop the formation of an irrational belief in the other' (1984: 86). So basically we can see that the sub-system and the main system are playing the roles of deceiver and deceived respectively here, and their status as separate subjects of psychological attributions affords the means to overcome the dynamic paradox. Because we conceive of the self-deceiver as broken up into two systems, we no longer have to attribute the conspicuously incompatible attitudes to the one subject.

In this way the homuncular theory gives at least the semblance of a solution to the problem. But there are few approaches to self-deception that have been more heavily

criticized than this one (e.g. Johnston 1988, McLaughlin 1996, Poellner 2004). It will not be necessary to enumerate all the objections here, though we may run through a few. Many of the objections to Pears' theory highlight things that the homuncular theory just leaves unexplained and mysterious. Johnston, for instance, poses a number of questions that remain unanswered in Pears account. According to Johnston, Pears has nothing enlightening to say about how the sub-system installs the belief in the main system. It is also left unexplained how the main-system is so susceptible to manipulation by the sub-system. Furthermore, we are left scratching our heads over why the sub-system is so concerned with the main system, and if it is so concerned, why it sometimes installs illusory beliefs which are not in the main-system's long-term interest.

A more fundamental objection to this approach, however, would question the very intelligibility of Pears' practice of ascribing psychological predicates that are applicable to the person, to something less than the person (a part of the person or something inside the person). Bennett and Hacker (2003: chap.3) have called this the *mereological fallacy*: the fallacy of ascribing predicates to parts that are supposed to apply only to the whole. Here the idea is that it is the whole (i.e. the person or human animal) that displays the behaviour that serves as the criterion warranting the ascription of the psychological predicates (belief, consciousness, intention, desire etc.), but not the whole's parts. As Alexander Bird puts it when criticizing homuncularism, '…the criteria for the correct ascription of these attributes such as belief make reference to the behaviour of the possessor, but such behaviour can be exhibited only by the complete human being' (Bird 1994: 22). I won't delve any further into this issue, however, since perhaps less taxing objections to homuncularism are to hand.

The simplest and perhaps most devastating objection to homuncular theories is one that has been voiced by a number of commentators, though most lucidly and forcefully by Brian McLaughlin. Since I could not better his exposition of it, I will quote him at length. In

the following ‘S1’ and ‘S2’ represent the subsystems of the self-deceiver, and S1 deceives S2.

> Homuncular approaches face a host of problems. I will mention only the most pressing ones. Why does the fact that S1 is a deceiver make the person of whom S1 is a subsystem a deceiver? Similarly, why does the fact that S2 is deceived make the person of whom S2 is a subsystem deceived? Subsystem S1 is a deceiver, but not deceived. Subsystem S2 is deceived, but not a deceiver. If the person is S1, then the person is not deceived, for S1 is not deceived. If the person is S2, then the person is deceived. But, then, the person is not a deceiver, for S2 is not a deceiver. Either way, the person is not both a deceiver and deceived. The person cannot, or course, be both S1 and S2, for S1 is not S2, and identity is transitive. Granting that the person is not identical with S1, could S1’s continuing to believe that *P* count as the person’s continuing to believe that *P*? And could S2’s ceasing to believe that *P* count as the person’s ceasing to believe that *P*? I don’t see how. For how could it be that the person both continues to believe that *P* and ceases to believe that *P*? If the person continues to believe that *P*, then the person did not cease to believe that *P*. Could S1’s performing an act of deception count as the person’s performing an act of deception? In a word, no. The notion of intentional agency transmitting from one agent to another is incoherent.
>
> I will stop, for homuncular approaches to self-deception seem to me to be nonstarters. Either states and acts of the subsystems count as states and acts of the person the subsystems constitute or they do not. If they count as states and acts of the person, then the puzzles of self-deception are re-introduced. If the states and acts of the subsystems do not count as states and acts of the person they constitute, then the

person is not in a state of self-induced deception and so is not self-deceived (1996: 41).

Another difficulty about the homuncular theory with respect to our case in particular, is that it is unclear how such a theory could be reconciled with the kind of empirical evidence about self-deception which we have been looking at. With respect to Basic Scenario cases, which are the exemplars of self-deception, we have seen that responsibility for self-deception can be attributed to the actions of the self-deceiver himself (he ends up self-deceived because of the way *he* evaluates the issue). That is to say, the *self-deceiver himself* engages in the culpable intentional actions. Nothing that we would want to call a 'sub-system' seems to be operating in these cases, and we do not seem to have any use for the sub-system/main system distinction when interpreting the data.

The embattled homuncular approach has few if any friends nowadays. Pears' mental sub-systems are a prime example of what Mele has called 'mental exotica' that are brought in to explain how intentional self-deception is possible in ordinary circumstances, the insinuation being that the explanation here is certainly no less and perhaps more philosophically puzzling/problematic than the explanandum itself. Instead, we should go back and review the initial assumptions which we are trying to accommodate with the introduction of the explanation, as opposed to swallowing a spider to catch the fly, as Pears does. For the reasons McLaughlin outlines, it is vital that any theory that may vindicate intentionalism about self-deception keeps to the idea that the culpable deceptive actions be undertaken by the *same subject* as is the one who gets self-deceived. And in the next chapter, we will look at a version of intentionalism, the most popular and promising version I would say, which does not violate this just constraint.

## *8. Concluding Remarks.*

I take it that if we are justified in thinking that doing something intentionally entails doing it knowingly, and I have argued that we are, then we are justified in rejecting the intentionalist interpretation of ordinary Basic Scenario cases. However, this certainly does not show that there is no scope of application for an intentionalist theory of self-deception at all. For the empirical evidence showed that certain kinds of actions were behind the formation of the biased belief in these cases, and we then reasoned that *these* biased actions could not be performed intentionally while the biased belief is produced. But for all we yet know, there may be other kinds of strategies of intentional self-deception which one could carry out knowingly and successfully, and as I will show in the next chapter, there is indeed one strategy in particular which seems more promising than the ones we looked at here, for which the idea of intentionally deceiving oneself using this strategy can't be straightforwardly dismissed as paradoxical. The intentionalist may then argue that this strategy may be undertaken in circumstances that have escaped the scrutiny of psychologists. In the next chapter I will elaborate on and assess this line of argument open to the intentionalist.

# Chapter 6: The Attentional Strategy

## *1. Introduction.*

Earlier, I argued for the knowledge condition on intentional actions, which is the view that doing something intentionally implies doing it knowingly. If intentional self-deception were only possible on the assumption that this thesis were false–that we can carry out some task intentionally while being completely unaware of what we are doing–then arguably, intentionalists would be in a precarious and perhaps quite untenable position. However, one does not find that intentionalists place all their bets on that assumption. Indeed, when one looks at the ideas intentionalists have about how intentional self-deception gets accomplished, one sees that in most cases they don't advert to unconscious intentional actions at all. They typically conceive of the self-deceptive belief as being brought about through a particular strategy, one which they seem to assume could be carried out consciously or knowingly, without this proving fatal to the self-deceptive project. Or at least such an assumption on their part might be inferred from the fact that they generally are not at pains to add caveats that this strategy to deceive oneself would have to be executed intentionally yet unconsciously.

In this chapter I will explain what this strategy is and how it is supposed to work. Then I will show how the constraint of the knowledge condition does not imply that it would be impossible for this strategy to succeed in the way it seemed to for the other candidate strategies we looked at, which is essentially tied up with the idea that this is a knowledge undermining strategy. After that, I will show how the intentionalist may meet the objection that no evidence has been found by psychologists that such a strategy is involved in self-deception, which they may do by claiming that it most likely occurs in more extreme situations which psychologists have not been able to study experimentally. Because the idea

of one succeeding in intentionally deceiving oneself using this strategy is not paradoxical, I will take it to offer the best prospects for an intentionalist theory of self-deception, and hence as a theory that must be taken seriously by us. However, I will then show that the idea that this strategy could work is opposed by a certain body of empirical evidence: the thought-suppression literature.

## *2. The Attentional Account of Self-Deception.*

Recently, Robert Lockie has usefully categorized a type of account of self-deception under the label 'attentional accounts', or what we may just call 'attentionalism'. According to this, self-deception is 'a phenomenon of attention: one attends to data conducive to the false conclusion, not-*p*, and does not attend to data conducive to *p*. In this way one deceives oneself into believing not-*p*' (2003: 131). The sense of 'deceives oneself' in use here is the 'strong' sense (Pugmire 1969) of *making oneself acquire a belief which, before executing the deceptive strategy, one believed to be unwarranted*. We may formulate how on the attentionalist account, the intentionalist sees self-deception as happening more precisely as follows:

> A person who after encountering unwelcome considerations (call them '*C*-') and forming the warranted, true, and unwelcome belief on their basis that not-*p*, but who wishes to believe that *p* (to avoid the distress of knowing that not-*p*), by turning his attention away from *C*-, and by finding and attending to more welcome considerations (*C*+) supportive of the contrary *p*, may cause himself to lose the belief that not-*p* and acquire the belief that *p*.

Some philosophers implicate similar attentional maneuvers in some 'weaker' cases that they classify as deceiving oneself, as with starting out knowing that *p* and making yourself loose that belief, without gaining the belief that not-*p* (Audi 1982. Whisner 1998)[40]. Starting out thinking *p* and not-*p* just as likely, and making yourself believe not-*p* could be another example of a weaker case. But let's narrow our focus on the most commonly discussed type of case here.

Note that by 'turning attention away' from the unwelcome evidence what these theorists primarily have in mind are not overt, physical acts such as, say, directing one's gaze away from evidence, but, more importantly, the deliberate avoidance of certain thoughts or the ridding of thoughts from the mind by the control of attention–what psychologists call *thought-suppression*–as well as selective focusing of attention onto more welcome considerations. For physically avoiding unwelcome evidence would not suffice for such a self-deceiver so long as she remembered or remained mindful of the existence and import of that evidence.

A quick illustration of this phenomenon may be in order. Say that Burke is a goal-keeper who has lately been letting in some easy shots. He's beginning to think that he's not a very good goal-keeper and his ego can't tolerate this realization. The attentional theory states that by shunning the thoughts of his poor performances and by concentrating on the few memories of when he performed well, Burke may come to subvert his belief that he is a bad goal-keeper and acquire the more welcome, contrary belief.

Let me now present some textual evidence showing how frequently these attentional shifts are implicated in self-deception. In Davidson's opinion, '[t]he action involved [in self-deception] may be no more than an intentional directing of attention away from the evidence in favour of *p*; or it may involve the active search for evidence against *p*' (2004/1986: 208).

[40] Though if the result of the action is not that the subject has any positively mistaken view, perhaps we should take this as a case of the repression of a belief rather than as a case of self-deception.

'An important way to become self-deceived' Perring urges, 'is by intentionally ignoring and avoiding the contemplation of evidence one has for an upsetting conclusion, knowing full well that one is giving priority to one's present peace of mind over the search for truth' (1997: 123). Van Leeuwen states that 'discomfort is there especially when my attention is on the evidence that ~*p*, but abates when my attention shifts–especially when it shifts to evidence that *p*… Self-deception ensues, for the focus of attention on the evidence that *p* gives rise to the *p*-belief' (2008: 198-199). For Whisner, one can lose a warranted belief 'either by shifting attention from a focused attention on the "object" [i.e. the ominous evidence] or by refusing to focus attention on the "object"' (1998: 204). Hamlyn agrees that a man 'can concentrate his attention on certain things to the exclusion of others' and thereby deceive himself (1971: 58). For Audi, an unwelcome belief can be made 'inaccessible', and '[p]resumably, the process often begins with *S*'s consciously putting the evidence against *p* out of mind, attending eagerly to any evidence there is for *p*, and the like' (1976: 383. Also see Audi 1982: 145). Furthermore, in Noordhof's view '[t]he distinctive feature of the self-deceived is that a failure of attentive consciousness enables them to possess or retain a belief that they would not otherwise have' (2003: 87), where this failure is due to a 'systematic avoidance' (89) of evidence against the favoured belief and of one's own avoidance strategies. Let's call theorists such as these 'attentionalists'.

All the above philosophers accept that *intentionally* deceiving oneself, in the sense of making oneself believe something one justifiably believed to be false or unwarranted, can be accomplished through such attentional maneuvers (though Whisner's opinion is not clear on this point)[41]. In fact, when intentionalists discuss what strategy people use to intentionally deceive themselves, they typically mention such behaviour. Moreover, none of these

[41] Though it's worth noting that, as I've said earlier, Whisner and Audi have weaker conceptions of self-deception where the result of these shifts is generally the loss or rendering unconscious of an unwelcome belief without the gaining of the contrary belief.

philosophers make any special effort to mention that these intentional actions are carried out unconsciously, and some are even explicit in saying the self-deceiver knows full well what he/she's doing (see again the Perring and Audi quotes above). Let us now examine how exactly these shifts of attention are supposed to result in these doxastic changes.

### *3. How is Suppression Supposed to Result in Belief Change?*

The presupposition of the attentionalist account is that changes to what we believe can occur not just in the usual way as when we encounter new evidentially relevant considerations etc., but in a deviant way, simply by shifting our thoughts and attention between the belief-relevant considerations we are already acquainted with. But we might wonder how this could ever be effective. The worry is well put by David Kipp:

> …the main difficulty with attention-oriented theories is that they make plausible merely how a person might contrive to avoid thinking about an unwelcome belief… To put something one knows out of one's immediate consciousness by directing one's attention elsewhere can effect, so long as successful, only that one ceases to think about that thing, and not that one ceases to know it (1980: 310).

So merely directing one's attention away from *C*- would not, in Kipp's view, make one no longer know about that evidence. It might relieve one of the anxieties associated with thinking about the unwelcome truth, for a while at least, but it would not relieve oneself of the belief in that truth itself.

However, many attentionalists seem to think that these attentional shifts can accomplish more than merely removing thoughts of the unwelcome evidence from one's

consciousness. Perring says that '[i]ntentional self-deceivers manage to maintain the belief that not-*p* by intentionally ignoring, blocking out, and engaging in activities *that will lead them to forget* evidence that *p*… they find ways to *forget* that information' (1997: 124, emphasis added). Whisner also writes 'On my account [the self-deceiver] might avoid the thought [of the unwelcome evidence or belief]…and if the thought avoidance is successful (he may have to engage in thought avoidance repeatedly) he *might succeed in forgetting* what he believed and become ignorant of what he believed and said because of his successful attempt to deceive himself' (1998: 196, emphasis added). Additionally, in Van Leeuwen's view, '[s]uppression may be used deliberately in willful self-deception to *undermine memory* of evidence contrary to the desired belief' (2008: 202, emphasis added). Hamlyn, who also claims that intentional self-deception can be accomplished through the attentional strategy, says that '[b]y putting something out of mind repeatedly a man may come as a result to forget that very thing' (1971: 46). These philosophers mention the issue of forgetting quite offhandedly, but it may be a crucial element of the strategy if that strategy is to do the trick. For if you believe that not-*p* because of considerations *C*-, then we ought to distinguish between making yourself *not think about C*-, and making yourself *forget about C*-, and it may be the latter that is necessary if one is to rid oneself of the belief.

The assumption of the attentionalist, then, must be that we can lose a belief by forgetting the considerations that were our reasons for believing it. If this assumption were false, the attentionalist would be in trouble, but it is an assumption I am willing to grant, since it seems plausible. The following, for instance, seems like an only too familiar occurrence. Jones has always voted for the People's Party. During their latest term in office, however, some members of this party behaved reprehensibly, broke their election promises, and made some foolish decisions, earning Jones' contempt in the process. He finally came to believe that the People's Party are a loathsome lot, and undeserving of his support in the next

election, due in 3 years. However, being an unfortunately typical voter, Jones' memory for political matters is not so good. He soon forgets his reasons for disliking the People's Party, and gets seduced by their good behaviour and wonderful new election promises in the year before the election. Because of his forgetfulness, he ends up with the belief that the People's Party are not such a bad lot, and are deserving of his vote.

Of course, in this story I have simply stipulated that Jones' forgetting of the facts that were the reasons for his belief led to him losing that belief. But the point is that this story is not an implausible one for that. It is the kind of thing that we know can easily happen. There has also been some experimental work done that shows similar links between memory and belief. For instance, peoples' current beliefs regarding their possession of certain traits have been shown to be determined by whatever memories about their own past behaviour happen to be accessible to them at that time (Kunda *et al* 1993: 65). It is therefore not altogether implausible to think that in principle, if shifting our attention off *C*- could lead to our forgetting *C*- as many attentionalists profess, then where *C*- are our reasons for believing that not-*p*, this strategy could effectively undermine the belief that not-*p*.

It is also worth noting that mention of such attentional maneuvers can also be found in rival, non-intentionalist accounts, as with that of the intentionalist's nemesis, Alfred Mele, who claims that we can end up *un*intentionally deceiving ourselves through such behaviour. It is worth looking at Mele's version of these events to appreciate how attentionalism cuts across many theoretical divisions in the self-deception debate.

### *4. Mele's Attentionalism.*

Mele agrees that self-deception can result from intentionally shifting attention between belief-relevant considerations, but with an important qualification. He illustrates with the

fictional case of young Beth. Beth's recently deceased father used to ignore her to play with her brothers. The memories of these occasions are upsetting to her, presumably because they make her believe she was the less favoured child[42]. Mele says that she 'may intentionally focus her attention on the pleasant memories [of her father showering her with attention]…and intentionally turn her attention away from [these unpleasant memories]', and then, '[a]s a consequence of such intentional activities, she may acquire a false, unwarranted belief that her father cared more deeply for her than for anyone else' (1997a: 98). However, as Mele describes the case, Beth's intention in shifting her attention in these ways was *not* to acquire that belief and to lose the upsetting one (and so was not an intention to deceive herself). Rather, her intention was the more humdrum one to *avoid dwelling on unpleasant memories*, and to *think of more pleasant matters*. Her acquiring the belief that her father loved her most occurred as an *unintended side-effect* of those intentional acts. She *un*intentionally undermined her belief/suspicion by undermining the memories that evidentially supported it (and for Mele, this would still count as self-deception). The intention with which she diverted her attention did not fall under more dramatic descriptions, such as 'wanting to change my belief, regardless of the truth'. So we can see that in different accounts, it is claimed that the same attentional avoidance behaviour can result in the same doxastic changes, though it can be done with different intentions, i.e. with the intention to produce those doxastic changes, or with other more humdrum intentions.

Mele has in a number of places expressed reservations if not outright skepticism about whether intentionally deceiving oneself is a realistic prospect (e.g. 1997a: 92), but the intentionalist might argue that it is a short step from accepting the possibility of cases like

[42] Mele does not explicitly say that Beth starts out with the belief/suspicion that she was less favoured. But it is difficult to explain her wanting to avoid thinking about those occasions, and to explain the fact that those memories upset her, without supposing that she realized their significance (i.e. that they mean she was less favoured). We can't explain these facts without supposing that she had at least an inkling of the unwelcome truth that the memories are evidence for. On this, see (Lockie 2003).

Beth's to accepting that intentional self-deception using the attentional strategy is a possibility. Mele accepts the basic attentionalist thesis–popular amongst intentionalists–that doxastic changes can be induced if we shift our attention in appropriate ways between belief-relevant considerations, though he does not think that these doxastic changes would have been intended in those cases. But although he says that doxastic changes can occur as an unintended side-effect of these shifts, the intentionalist might argue that if someone were to realize that such shifts can induce such changes, she might then exploit this knowledge to deliberately induce them. It would just be a case of the following:

1) *S* does *x* with the intention of achieving *y*.
2) In the process of doing *x*, *z* is accidentally brought about.
3) *S* eventually realizes that *z* can be brought about by doing *x*.
4) *S* starts doing *x* with the new intention of bringing about *z*.

In other words, the only difference between the Beth case as Mele describes it and as an intentionalist would describe it is in what the subject intends to achieve in behaving in that way; the behaviour is the very same in both cases. Mele may want to argue that the content of the intention would make all the difference here to whether the doxastic changes could occur, but it's not altogether clear, to me at least, how this is so. We should look on the act of trying to forget something as similar to the act of trying to sleep; in both cases we do things that *facilitate* a certain result happening, one which happens, once appropriate conditions are set up, of its own accord. In one case, one shifts attention to something else, in the other, one lies down comfortably and closes one's eyes. One can lie down comfortably and close one's eyes with the intention of sleeping, or only of resting. But no matter what one's intention is in doing it, so long as one sets up the appropriate conditions, sleep is likely to follow. With

thought-suppression, the condition for losing consciousness of *x* is that your attention is trained on *y*. What your ultimate intention was in shifting your attention to *y* should not matter. (In the next section we will examine the nature of the action of trying to forget in greater detail).

Let us now compose a more detailed definition of the attentionalist theory, incorporating all that has been said. Note that with this definition, I intend 'attentionalism' to include *both intentionalist and non-intentionalist Beth-style cases* (since ultimately, I think that the theses that either of these scenarios are possible will stand or fall together). Note that, as Whisner and Hamlyn suggest, on this view the self-deception may take time. It may involve repeated attempts at suppressing *C*-, thus rendering *C*- progressively less accessible to recall, finally resulting in the loss or 'burial' of the unwelcome belief.

> Self-deception can occur when a person, who after encountering unwelcome considerations *C*- which make her believe that a distressing proposition not-*p* is true, intentionally shifts her attention from *C*-, and onto considerations *C*+ which are more supportive of the contrary, agreeable proposition *p*. As a result (perhaps after repeated attempts), she ends up forgetting *C*-, and consequently she loses the belief which *C*- supports. Because now only *C*+ is all she remembers, the considerations she has to judge from weigh in favour of *p*, and she acquires that belief. Her attentional shifts may have been done with the intention of causing these doxastic changes, or with the simpler aim of not thinking about unpleasant matters, where these doxastic changes occurred as an unintended side-effect.

*5. Why the Attentionalist Theory can Overcome the Problem of the Knowledge Condition.*

As I have said, intentionalists who assume that the self-deceiver uses these attentional shifts to deceive himself (we may call these attentionalist intentionalists, or just A-intentionalists for short) are not particularly insistent that these shifts are done with an *unconscious* intention to deceive oneself. The assumption seems to be that one could be aware of what one is doing here, without this making it impossible to accomplish the deception. If it were true that one didn't *need* to suppose unconscious intentional activity here, this would indeed be an advantage of the attentional account, since as we saw, it is far from certain that there is a legitimate notion of unconscious intentional action to draw upon. But why would the attentional strategy be any different from the other strategies we looked at with respect to the knowledge condition? Why wouldn't the knowledge condition prove just as fatal to that strategy as it seemed to do for the others? The answer, I believe, has to do with the fact that *the attentional strategy is a knowledge suppressing strategy*. For the other strategies usually mentioned, the knowledge condition would prove fatal because nothing was done to undermine the unwelcome knowledge that gets in the way of one having the welcome belief. But the attentional strategy is *specifically geared towards suppressing any unwelcome knowledge which may stand in the way of our having a belief*. That is what sets it apart from the rest.

If we remember, for instance, our examination of the idea of intentionally rationalizing, which we defined as intentionally adducing specious arguments, in this case we assumed that granting one does this with the knowledge of what one is doing, one remains with the knowledge that one has just produced a specious argument. This knowledge, then, seemed to be an obstacle to one's being swayed by that argument. But the attentional strategy is specifically geared towards suppressing any knowledge that would be an obstacle to our

having the welcome belief, by shifting attention away from that knowledge and keeping it out of mind. This is why the attentional strategy seems to stand out as offering us the prospect that intentionally deceiving oneself is possible, even granting the knowledge condition.

But does it? For is it not being said here that the event of forgetting about or ceasing to thinking about the evidence is something the thought-suppresser does intentionally? And then wouldn't this imply that she is doing it knowingly, such that she knows that she has forgotten it? And paradoxically, this seems to imply that the evidence is in mind, so that it has not really been forgotten after all. Pugmire (1969: 346) and Reilly (1976: 393) voice worries similar to this.

In fact, it is not being said that the self-deceiver intentionally forgets, since forgetting isn't, and couldn't be, an intentional action (any more than falling asleep is). Rather we are saying that she intentionally *makes herself* forget. The event of forgetting in these cases is not an intentional action but is supposed to be the intended result of other things that we do intentionally. Here it will be useful to look at what thought-suppression typically involves. Daniel Wegner–one of the most prominent researchers on this issue–says that when people try to take their mind off something, they typically do so by putting it onto something else (1994: 12 & 60). They turn their mind or attention to a 'distracter' which may absorb their attention, thus keeping it off the unwanted thought.

We can then look at thought-suppression as an act of the same sort as going to sleep. In both cases, we intentionally do things to *facilitate* the occurrence of certain events, i.e., certain thoughts being lost from consciousness, or our losing consciousness altogether. We may facilitate these results in the one case by lying down and shutting our eyes in a dark, quiet room, and in the other, by training our attention onto something else. These actions are aimed at exploiting certain mental or psycho-physical regularities which occur independently of our will (i.e. the regularity between being in a certain psycho-physical state and falling

asleep, and having your attention trained on *y* and forgetting about *x*). And once appropriate conditions are set up, the regularity takes its course and no further conscious monitoring is required.

Thus we should understand the bringing about of these mental events as instances of what Matt Soteriou, following Fabian Dorsch, calls 'mediated agency' in bringing about a mental event, when 'we "trigger" some process (epistemic or merely causal) with some goal in mind, but recognize, and instrumentally rely on, the capacity of such a process to lead, *by itself*, to the desired outcome' (Soteriou 2009: 5). Or following Mark Johnston, we would speak of the utilization of 'autonomous means' in the acts of trying to suppress a thought, by which he means a mental process or regularity that, once initiated, takes care of itself (Johnston 1988). So the fact that the self-deceiver intentionally/knowingly tries to suppress a thought doesn't imply that when the thought is forgotten, she knows that it is, or that she is intentionally/knowingly not thinking about it, any more than our intentionally/knowingly going to sleep implies that when we finally fall asleep, we know (or are aware) that we are asleep, or that we are intentionally sleeping.[43] These results are things that we intended without being things we are doing intentionally, and may be achieved through the exploitation of a mental or psycho-physical regularity which occurs independently of the will.

The idea, then, is that when the unwelcome evidence is in mind, there is a moment where we shift our attention away from it 'knowing full well that one is giving priority to one's present peace of mind over the search for truth' (Perring 1997: 123). But then, after forcing our attention elsewhere, we forget about the evidence as well as about the fact that we just deliberately tried to forget it. By such means we can make ourselves loose the belief that the knowledge of such evidence supports.

[43] Suicide is another example. One commits suicide intentionally, but this doesn't imply that one knows that one has succeeded when one has. One just takes some pills or pulls a trigger and thus initiates an 'autonomous means'.

The idea that one might successfully try to deceive oneself into believing something that one knew is false using the attentional strategy is, I believe, not a paradoxical idea, for the reasons given. The attentionalist picture of intentional self-deception is thus a coherent one. However, it looks like the A-intentionalist is faced with a problem in the form of the experiments we looked at in chapter 2. These experiments concerned situations where subjects encountered evidence for an unwelcome proposition, and the experimenters tested for whether people respond in such situations by shifting and keeping attention off this unwelcome evidence. And as we saw, it was found that subjects didn't do this, but did the exact opposite; they attended to that unwelcome evidence in the attempt to refute it. They took a 'fight' response instead of a 'flight' response. This may be a problem for the A-intentionalist because though it seems as though cases of the sort he describes are logically possible, we could say on the basis of those experiments that the evidence suggests that people, as a matter of fact, don't react in this way when they encounter unwelcome evidence. They react, instead, with rationalization, a kind of behaviour which is not, as we saw, amenable to an intentionalist treatment.

However, there may be room for the A-intentionalist to find space for her theory. She may argue that although psychologists have not found evidence for the behaviours described in the attentionalist theory, in more *extreme* cases, which may have eluded the laboratory of the experimental psychologist, such strategies may be operative. Indeed, this seems to have been the strategy that Christian Perring has taken to defending A-intentionalism. This issue will be discussed in the next section.

*6. Finding a Place for the Attentional Account.*

To mention again, the empirical studies into Basic Scenario cases that we looked at in chapter 2 did *not* find behaviours such as the attentionalist theory describes to be behind the formation of motivationally biased beliefs. In these studies, the hypothesis that such shifts of attention away from the unwelcome evidence are operative was tested for by giving subjects a memory test to see if they remembered the details of the unwelcome evidence after they were confronted with the relevant information (see Liberman & Chaiken 1992. Wyer and Frey 1983). Because subjects knew the details of the unwelcome evidence quite well afterwards, it was concluded that they were not trying to ignore it. Rather, they attempted to refute that evidence, which requires *paying* attention to it. So is the idea that the attentional strategy is operative in cases of self-deception a myth of some sort?

The A-intentionalist may respond no, arguing that these experiments don't prove that these attentionalist cases don't exist at all. Here she may not simply be referring to the idea that 'you can't prove a negative'. For it may be that the cases which are most plausibly explicable within an A-intentionalist framework are the more *extreme* ones that can't be replicated in the lab. Christian Perring makes this defense of his attentionalist version of intentionalism against the kind of anti-intentionalist approach that relies on the results of experimental psychology. He writes:

> I conjecture that intentional self-deception occurs mostly when a person gains access to very upsetting evidence. The less emotionally charged experiments that Mele does discuss...are more likely to be explicable just by his motivated mistakes model. If this is right, then there is a specific problem in measuring intentional self-deception. We cannot create such experimental situations in psychological laboratories, with

> volunteer subjects being given hurtful information, if only because of the ethical restrictions placed on researchers (1997: 124).

It is difficult to reproach Perring here because as he points out, it is not easy to get the opportunity to scientifically study how people react when they are confronted with greatly distressing evidence. The non-intentionalist may reply that there's no reason to think that it would be impossible for the kind of processes we have looked at to be capable of causing the acquisition/maintenance of unwarranted beliefs in cases where the issue is much more worrying than the issues subjects are presented with in the average psychology experiment pertaining to self-deception. Nevertheless, that's hardly a knock-out argument against the A-intentionalist.

Further space may be opened up for the A-intentionalist account as follows. The psychological studies into the influence of desire on belief which we have looked at – studies which, as Perring suggests, are best explained in non-intentionalist terms – suggest that these unintended motivated biases in one's judgment only occur where the evidence is *not conclusive*. This is something that is often not appreciated in the non-intentionalist philosophical literature, where there can be a presumption that our powers for self-deception are nothing short of extraordinary. Ariela Lazar, for instance, asks, '[h]ow does a subject who is competent to detect the irrationality of a belief that *p*, form her belief against *weighty* or even *conclusive* evidence to the contrary?' (1999: 265, my emphasis). Her view is that '[t]he self-deceived subject is highly irrational–she forms a belief that is undermined by the weight of the evidence even to the extent of the evidence being *overwhelming* in support of its negation' (1999: 268, my emphasis).

But there is a discrepancy between these kinds of remarks, and the more restrained remarks of psychologists who study empirically the phenomenon of motivationally biased

belief. Ziva Kunda cautions us that, '…an explanation for how directional goals affect reasoning has to account not only for the existence of motivated biases but also for the findings suggesting that such biases are not unconstrained: People do not seem to be at liberty to conclude whatever they want to conclude merely because they want to' (1990: 482).

We may say that motivationally biased belief can only occur when there is *scope* for it to occur. The kinds of factors that would be mentioned by psychologists as creating scope for this to occur would include the *lack of conclusive considerations supporting the truth*, and the availability of some argumentative material–however specious–for making a case or constructing an apparent justification for the agreeable position. As E.A. Johnson observes, '…self-deception constitutes a kind of shadow-land epistemic phenomenon that comes into being only under cover of ambiguity or uncertainty, and evaporates when exposed to the light of overwhelming fact' (1997: 118. Also see Baumeister 1993: 174). For instance, biased reasoning occurs more easily when the subject is faced with ambiguous data open to different interpretations, and not when the data is unequivocal (Dunning 1999: 3-4), and it is also facilitated when there is an available stock of inconsistent information pertaining to an issue rather than a stock that consistently suggests a particular conclusion (Kunda *et al* 1993: experiment 2. Sanitioso *et al* 1990: 239). These points are well put by Béla Szabados:

> Philosophers are apt to say the evidence against the proposition that a person deceives himself into believing must be 'overwhelming', *must* be 'strong,' or that the evidence *must* obviously or plainly be against the proposition that the self-deceiver believes. This is misleading. For the evidence in question obviously must allow for the

> possibility of adjustment, colouring, manipulating, interpreting–otherwise, the very possibility of self-deception is ruled out (Szabados 1974: 66[44]).

Part of the point here is that in these cases of motivationally biased belief, the person makes his belief out to have a *semblance of rationality*. To do this he feels the need to *construct a justification* for the desirable view. For this to be possible, the evidence for the undesirable conclusion must *not* be incontrovertible, but must allow scope for reinterpretation and rationalization (explaining away). As stated in the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders*, 4th edition, false belief that is maintained in the face of 'incontrovertible and obvious proof or evidence to the contrary' is the mark of *delusion* (2000: 821), rather than self-deception.

How these points make room for A-intentionalism is as follows. Perhaps, as we have said, intentional self-deception accomplished by means of the attentionalist strategy is something that happens in more extreme scenarios. But these cases may be 'extreme', not just with respect to the degree to which the unwelcome evidence is distressing, but also with respect to the degree of strength of that evidence. That is, this phenomenon may be more liable to occur in cases where the subject encounters conclusive or near conclusive evidence that an extremely alarming proposition is true. In such cases – notwithstanding the subject's tendency towards motivationally biased reasoning – it may be simply impossible to avoid drawing the unwelcome conclusion, and such may be the distress that this puts her under, that she may be desperate to escape from this realization. Our natural concern to know the truth may in such circumstances get completely overridden by the desire to protect ourselves from

[44] Another pertinent remark by Szabados is: '[Take] such beliefs as "I am a human being", "I have a body", "I have sensations". With regard to these beliefs there seem to be no room for explaining away the evidence, manipulating the evidence, distorting the evidence. It seems to me that there is no room here for self-deception either. Anyone who believes that he is not a human being, that, say, he is made of glass; that, say, he is a pumpkin, is not, it seems to me, self-deceived but deranged, mad' (1974: 66).

anxiety, and the subject may put into operation the attentional strategy. However, had the evidence warranted belief in the alarming conclusion without proving it conclusively, then perhaps the subject would have not drawn the unwelcome conclusion because she would have irrationally discounted and explained away that evidence by reasoning about the issue in a motivationally biased fashion. She thus would not have needed to execute the attentional strategy.

This suggestion cannot be undermined by adverting to laboratory studies, since such studies do not recreate these kinds of circumstances. Subjects in psychological studies into the influence of desire on belief formation are never presented with incontrovertible evidence for the truth of an extremely unsettling proposition, if for no other reason, as Perring said, than because of the ethical restrictions on such research. So the reason why no evidence for the existence of the attentional strategy has been found, it will be argued, may be that we haven't been looking in the right places for it. Intentional self-deception accomplished through the attentional strategy, it will be insisted, may very well be a reality. There may be explananda for which the theory of intentionalism thus has application, to be recognized alongside explananda for which a non-intentionalist approach is more appropriate.

However, I think that for a non-intentionalist to concede even that there may be a certain domain where intentionalism has application would be to concede too much to her opponent. For there is a certain body of empirical evidence that stands in opposition to the claim that one could successfully deceive oneself through the attentional strategy, be it in the 'extreme' circumstances we have just mentioned, or any other. What I have in mind here is the literature on thought-suppression. Thought-suppression, or the act of trying to keep an unwanted thought out of one's mind, is an essential component of the attentionalist theory, according to which one can make oneself forget certain things which one knows by suppressing from consciousness one's thoughts or memories of them. The attentional theory

seems to be premised on the idea that it is possible to do this, and I have argued that it is indeed *logically* possible that one could do this (i.e. the idea of one successfully doing this involves no obvious logical paradox, contradiction, or incoherence). However, the thought-suppression literature suggests that it would be *psychologically* impossible for us to successfully repress our knowledge by means of thought-suppression. This is what will be discussed in the next section.

### *7. Can Suppression Result in Repression?*

David Kipp (1980), who discussed the idea that people deceive themselves by the control of attention, anticipated the point that these shifts are supposed to undermine the unwelcome belief by leading us to forget what we know. It was an idea he found psychologically unrealistic (also see Martin 1998: 191-192). But this view is drowned out by the more prevalent assumption that such a thing can happen. At this point philosophical debate reaches its limit, and the question of whether self-deception, on the attentionalist account, is possible comes within the jurisdiction of psychology. For it is, apparently, an empirical question whether shifting your attention off *C*-, repeatedly perhaps, can lead to you forgetting *C*-. Unfortunately, attentionalists have overlooked the relevant empirical work on this issue. To this we will now turn.

There have by now been many studies done investigating intentional thought-suppression, and some of these have studied the effects of suppression on accessibility of the suppressed item to later memory-recall. Some experiments tested the effect of suppressing affectively-neutral memories on their accessibility for recall and found modest reductions in

their accessibility (Anderson & Green 2001).[45] However, later attempts to replicate these results failed, finding suppression to have no effect on later recall ability (Bulevich *et al* 2006).[46] Of more relevance to the self-deception debate, however, are studies that focused on distressing thoughts. Rassin *et al* (1997) conducted an experiment where subjects watched a disturbing film fragment of a man being attacked by a bear. They were sent home, but had to return to the lab five hours later, with half of them instructed to suppress thoughts of the clip in the meanwhile. On their return, they completed a memory test on the details of the clip. Suppressers' memory was found to be no worse than that of non-suppressers.[47]

In fact, not only has little evidence been found that suppression can lead to forgetting to any significant degree, but research has highlighted the difficulty involved in simply keeping our minds off intrusive or unwanted thoughts. As Wegner and co-workers remark, the thought-suppression literature 'indicates that people have trouble even eliminating a thought from consciousness, much less erasing it from memory' (Wegner *et al* 1996: 680). The most remarkable finding here is probably what is termed the 'ironic' or 'paradoxical effect' of thought suppression: the so-called 'rebound effect'. This refers to the fact that the attempt to avoid thoughts often makes those thoughts intrude on one's mind even more. For instance, in one experiment which tested the long-term effect of suppression on thought-

[45] Van Leeuwen–one of the few philosophers that have referred to the empirical work on thought-suppression–claims that self-deceivers employ repressive strategies to achieve their end, and cites this experiment as 'put[ting] to rest' (2008: 202) the issue of whether repression is possible. It is unfortunate that he just referenced this study, since it is unrepresentative of the field as a whole.

[46] These experiments consisted of three stages. In the first, subjects had to learn unrelated word-pairings (e.g. ordeal-roach). In the second, they were presented with the first 'cue' words on a screen, and had to either call out the associated word, or in other cases, suppress the thought of it. In the third stage, to test the effect of suppression on recall, subjects were given the cue words (ordeal-?) and had to recall the associate, whether previously instructed to suppress it or not. Or they would be given an associated word with the first letter of the target word (insect-r_) and had to complete the word.

[47] Suppressers also had more thoughts about the clip over the five hour period, manifesting suppressions 'ironic effect', which will now be discussed.

frequency over a seven day period, subjects were instructed to think for two minutes about the most distressing event they had experienced during the last two years, and then to keep this memory suppressed as best they could for two minutes. It was found that the subjects who were most successful at rebuffing these thoughts during this time were the ones that suffered the most from intrusions of this thought over the following seven days, thus demonstrating the 'rebound effect' over a long-term period (Geraerts *et al* 2006). The rebound effect has been found to be a quite robust phenomenon (see Rassin 2005). '[I]t is safe to conclude' writes Rassin in a literature review, 'that suppression is ineffective in the short run, and even has a paradoxical rebound effect in the longer run' (2005: 54). Of course, the fact that unwanted thoughts so easily return to mind after their suppression implies that they haven't been forgotten. But it is not just that suppressing thoughts does not lead to our forgetting them. In Wegner and Erber's words, such attempts often have the opposite effect, rendering them 'hyperaccessible' to consciousness (Wegner & Erber 1992).

Note that it seems more accurate to call the kind of impossibility at issue here a psychological one rather than a logical one. It seems conceivable that one could make or help oneself to forget something by directing attention away from it whenever it comes to mind. Otherwise why would psychologists test for this possibility experimentally?

But are these results really relevant to our attentionalist cases? Arguably they are. Firstly, in the attentionalist picture of self-deception, the subject tries to keep out of mind thoughts which have negative affective significance. The thoughts in some of these experiments were also of this character (as with trying to keep out of mind thoughts of the most distressing event that one had experienced in the last year). A critic might insist that the degree of anxiety evoked by the unwanted thoughts in the attentionalist cases is supposed to be greater than in these experimental cases. However, there seems to be no reason to think this would make them any easier to repress, and it is well known that greater affective

significance associated with a memory or item of knowledge is linked in generally with greater vividness, longevity, and accessibility of that memory/knowledge to consciousness (e.g. Walker *et al* 1997). The critic may state further that there is an element of desperation behind the attempts in the attentionalist cases which is missing in these experimental ones. But there is no reason to think that this should make the thought-suppressers in the attentionalist cases any more successful, for the point about the rebound effect seems to be that the more effort one puts into trying to suppress thoughts, the *less* successful one will be.

It could be objected that in these experimental situations, the subjects were explicitly instructed to suppress certain thoughts, and proceeded to carry out this instruction in a conscious, deliberate manner, while in the attentionalist cases it's more probable that these actions occur unconsciously. However, this position is not available if one wants to maintain that the self-deceiver does these actions intentionally, with the intention of deceiving himself. Or at least this is so if it is true, as was claimed earlier, that doing something intentionally implies doing it knowingly or consciously. I hope that this thesis was argued for adequately already. And it is not evident that there are any other important differences between what subject's are trying to do in these experiments, and what subjects are supposed to be doing in the attentionalist cases, which would preclude us from legitimately extrapolating from the former to the latter.

It is worth noting here that although our primary interest is the relevance these studies have for the intentionalist version of attentionalism, they are of equal relevance to the non-intentionalist version. With regard to Mele's Beth case, Beth was also engaging in attempts at thought-suppression in trying to keep her mind of unpleasant memories about her father by training her attention on more pleasant memories, and this evidence suggests that she probably wouldn't have succeeded very well in even in keeping her mind off those memories, much less in completely forgetting them.

To sum up, attentionalist philosophers should be aware that the empirical evidence is weak at best that we have the kind of mental control over ourselves which attentionalist accounts of self-deception are premised on. Notwithstanding the would-be intentional self-deceiver's efforts to ignore her knowledge of the alarming evidence for *p*, that knowledge will remain accessible to her consciousness, so that as soon as she considers the *p*-issue, that knowledge will be there to inform how she thinks about the matter, whether she likes it or not.[48] All in all, there is very little reason to think that doxastic changes can be effected by a person's attempts at thought-avoidance. So there is little reason to think that intentional self-deception, on the attentionalist account, is possible.

## *8. Concluding Remarks.*

To recap, we have looked at the difficulties facing intentionalist accounts of self-deception. We said that for intentional self-deception to be possible, there must be a feasible strategy one could use to intentionally deceive oneself. We established a general constraint on this search for a feasible strategy in the form of the knowledge condition on intentional action. We then saw that this constraint rendered the idea of intentionally deceiving yourself paradoxical with respect to a number of prospective strategies of intentional self-deception, including rationalization and selective evidence gathering. Because rationalization and selective evidence gathering were among the primary processes of self-deception operative in common Basic Scenario cases, this led to the conclusion that an intentionalist interpretation

[48] Indeed, this seems to be a crucial decider for whether *C*- has been forgotten or not: whether it comes to mind when she considers whether *p*. For this is an occasion when we should expect her to recall *C*- if *C*- hadn't been forgotten, since *C*- is of evidentially of crucial relevance to the *p*-issue. If *C*- didn't come to mind to influence her judgment when she considers whether *p*, then plausibly, *C*- has been forgotten.

of Basic Scenario cases is not a valid option, and these cases are best explained in non-intentionalist terms.

We then looked at the attentional strategy, which seemed to hold more promise in the search for a feasible strategy of intentional self-deception, since the idea of intentionally deceiving yourself using this strategy was not, I argued, a paradoxical idea, and we saw that the intentionalist had the option of arguing that this strategy may be operative in certain more extreme cases which have escaped the inquires of psychologists who study self-deception. However, from considering certain empirical literature, we saw that because of the way humans are psychologically constituted, the attentional strategy would not be a feasible strategy by which one could intentionally deceive oneself after all. The conclusion is, therefore, that there is no good reason to believe that intentional self-deception is something that's possible at all for ordinary people in ordinary circumstances (i.e. where 'special means' are not available). This concludes my discussion of the intentionalism versus non-intentionalism debate between traditionalists and deflationists.

# Part II: The Doxastic Condition of the Self-Deceiver

## Chapter 7: Contradictory Beliefs and How it is Possible to Have Them

### *1. Introduction.*

The next two chapters will primarily be about the contradictory beliefs feature which traditionalists associate with self-deception. Frequently, those who hold that self-deceivers have contradictory beliefs take this to be possible because one of the beliefs is unconscious (the unwelcome one that is). In this chapter, I aim to explicate the notion of unconscious belief, and to show how it is closely associated with the attentionalist account of self-deception discussed earlier. This will be done by showing that the outcome of a successful execution of the attentional strategy can be understood to involve the unwelcome belief being rendered unconscious. However, the empirical evidence on thought-suppression looked at earlier–insofar as it stands opposed to the idea that the attentional strategy could be successful–also stands opposed to the idea that one could render one's belief unconscious. Can we therefore conclude that self-deceivers don't have contradictory beliefs?

Although it seems that people do not have the power to render their own beliefs unconscious, there is a reply that the defender of the contradictory belief thesis could make. She may contend that it's not necessary for one of the beliefs to be unconscious in order to have contradictory beliefs. In particular, commentators on Davidson have often taken him to offer a view of the contradictory belief condition according to which neither of those beliefs are unconscious. The main goal of the rest of the chapter will be to show that this route for rescuing the contradictory belief feature is not open, because the only way one could have contradictory beliefs is if one of them was unconscious. I will also argue that–*contra* the

interpretation of these commentators–Davidson probably did hold that one of the beliefs must be unconscious if one is to have contradictory beliefs.

In the 8th and final chapter, I will discuss a phenomenon often associated with self-deception, referred to as its 'tension', which the contradictory belief hypothesis has been considered well suited or even necessary to explain, and which thus provides a large measure of the motivation for that view. It may therefore be thought that there is an explanatory lacuna left in my account if it gives up on this idea. In chapter 8, I will give an alternative way of explaining the tension of self-deception without supposing contradictory beliefs, by instead supposing the self-deceiver to have an unwarranted degree of conviction in the proposition, rather than an unwarranted belief.

## 2. *The Puzzle of Contradictory Beliefs.*

The idea that self-deceivers believe that *p* and simultaneously that not-*p* is an idea that has been looked on with suspicion. Some have even declared it impossible. Let us start, then, by asking the most basic question: is it possible to be in such a condition, and if so, how?

Consider expressions of the form '*p and not-p*', and '*S believes that p and that not-p*'. The first is an outright contradiction, and so does not express a possibility. The second, however, is the ascription of propositional attitudes to a subject and does not express a contradiction, though one might feel it to be problematic in some way. Nevertheless, *if* there is something problematic about interpreting someone as believing that *p* and that not-*p*, it will surely have something to do with the fact that that '*p* and not-*p*' is a contradiction and impossible.

Let us try out some proposals as to what a case of someone believing that *p* and that not-*p* would look like. Typically, people who believe that *p* are prepared to state that *p* is the

case and assent to it in thought. And people who believe that not-*p* are prepared to state that not-*p* is the case and assent to that in thought. So in that case, would someone who believes that *p* and also that not-*p* be someone who would be prepared to state '*p* and also not-*p* are the case', assenting to this conjunction in thought? John Campbell has recently argued that anyone who speaks in such a way that transgresses our basic logical norms could not really be interpreted as being as wildly irrational as their utterance might suggest (Campbell 2001). Campbell refers to Quine's view that if we were to come across a tribe of people who seemed to be transgressing our basic logical rules (to illustrate with reference to our case, if they appeared to be asserting '*p* gun da-*p*' where we translate this as '*p* and not-*p*') then we should conclude that we have mistranslated their utterances, rather than conclude that they are wildly illogical (that is, we should conclude that 'gun' and 'da' do not represent the logical operators of conjunction and negation, or at least that they don't in this syntactical/external circumstance, where they might be giving '*p* gun da-*p*' some special use). The reason, as I see it, has to do with the fact that the rules of logic don't just demarcate the logical from the illogical, but also *sense from nonsense*. If that is so, then to interpret someone as using language in a completely illogical way, is to interpret them as speaking nonsense, which is to interpret them as, strictly speaking, not really speaking at all. Therefore, if we are to hold onto the assumption that the subject is engaged in a genuine or meaningful linguistic activity, we must assume that they are conforming to the rules of logic, and interpret their utterances accordingly. Linguistic interpretation requires a presumption of logicality.

So a person saying that *p* and not-*p* could not be a good reason for interpreting them as 'really believing' *p* and not-*p*. Yet one might say that there are everyday cases of people believing contradictory propositions. R.J. Yanal speaks for the person putting forward this view in the following:

> People hold contradictory beliefs all the time. A Christian might, for example, believe that after death we continue to exist though as pure spirit, and also believe that after death, if we were wicked in life, we will suffer by being burned in fire. Pure spirit cannot by definition be physically burned, so this Christian's beliefs are contradictory. If self-deception is simply a case of someone holding inconsistent beliefs, self-deception isn't a unique doxastic category calling for some special explanation (2007: 113).

Let us assume with Yanal that these beliefs–ones which the Christian has avowed on separate occasions–are in fact contradictory. In such a case, the Christian's attention might be drawn to this fact. But it might happen that the Christian insists that there's no inconsistency. He might believe that the two beliefs do not contradict each other, and he might justify his case with some argument (which he incorrectly thinks is sound). Many cases of holding contradictory beliefs are like this, where the person holds two contradictory propositions as true, which he has considered together and is prepared to assert together, precisely *because* he doesn't believe them to be contradictory. However, such cases are only possible because the propositions don't *directly* contradict each other. The propositions 'after death we continue to exist though as pure spirit' and 'after death, if we were wicked in life, we will suffer by being burned in fire' do not stand to each other as *p* to not-*p*. That they contradict each other can be seen, if at all, by considering their implications and considering them in light of presupposed premises (such as the premise that one needs a physical body to get burned). Let's call this a case of *indirect* contradiction. The phenomena we are interested in, then, is one in which the subject believes propositions that *directly* contradict each other, as if the Christian were to believe that 'after death we continue to exist though as pure spirit' and also 'after death we do

not continue to exist though as pure spirit'.[49] But it seems impossible to imagine how anyone would believe *them* simultaneously *by* believing that they don't contradict each other, since we don't have to examine their implications or examine them in light of presupposed premises in order to see that they are contradictory. If someone were to deny that *p* and not-*p* are contradictory, we would again be forced to interpret his utterance 'non-literally', if we are to interpret it as being intelligible speech at all, or we would be forced to interpret him as not understanding our language.

So what we want to know is how it is possible for a subject to simultaneously have *directly* contradictory beliefs. In the next section I will present the most common answer to this, which employs the notion of an unconscious belief.

### *3. Unconscious Belief.*

The fact that one can have two directly contradictory beliefs (or knowledge directly contradicting belief[50]) has often been said to be possible if one of those beliefs is 'unconscious' (Hamlyn 1971: 57. Steffen 1986), 'inaccessible' (McLaughlin 1988: 51-53) or 'buried' (Haight 1985: 245. We may take these terms to be synonymous, though I will mention possible differences in meaning shortly). Skepticism as to the significance of the phrase 'unconscious belief', however, has been expressed by Mele, who doubts whether it means anything more than that the subject doesn't have the belief at all. He queries, '…if [the self-deceivers efforts] can be effective, one wonders why we should not suppose that, rather than rendering the knowledge (or belief) unconscious, it simply *eradicates* the true belief'

[49] Richard Foley describes this distinction as one between 'inconsistent' and 'contradictory' beliefs (1986: 328).

[50] Since the unwelcome belief that not-*p* for the self-deceiver is typically true and also warranted by the evidence that he was acquainted with, we might be entitled to speak of knowledge here as opposed to mere belief.

(1987: 4). In this section, I hope to clarify the idea of a belief being unconscious, and to show that it does mean something importantly different from saying the belief is gone.

Let's try and get a grip on the idea of how unconscious beliefs may be involved in self-deception with reference to the case of Beth that we discussed before. In this case, as Mele described it, Beth came to lose her belief that her father loved her the least in her family by turning her attention away from the memories which evidentially supported this belief, namely, memories of her father ignoring her and giving more attention to her brothers. Now it might be said that this would be a good example, not of someone getting rid of her belief, but of her *rendering it unconscious*. In choosing this expression we would be emphasizing the point that, though this belief won't come to mind, the *potential* for it to come back to mind has not been lost (i.e. though it doesn't come to mind when it would be expected to, it might do so in response to certain strong cues or reminders, or under certain conditions, like if someone were to describe to Beth the occasions she made herself forget). On the other hand, in saying that the belief is eradicated we are implying that this potential is not there (we also may be implying that the belief still influences thought and behaviour in certain subtle ways, though as I will discuss, I don't think this is a necessary condition on a belief being unconscious). This idea is implicit in Haight's elucidation of 'buriedness' as: 'I may have buried knowledge of some truth (T) when (i) I have at some time learned it, and (ii) nothing like a damaged brain now makes me forever unable to recall it, and yet (iii) I do not recall it, when recollection is to be expected' (1985: 245). The difference between a belief being gone and being unconscious would be analogous to the difference between being unable to find a book in a library because it has been misplaced on the wrong shelf, and being unable to find it because it was stolen and is not there anymore respectively.[51]

[51] Some philosophers have used the term 'unconscious belief' to talk about a belief that is not *being accessed* as opposed to one which is inaccessible (Mellor 1977/78, Searle 1992). On this view, my belief that my phone number is such-and-such is unconscious when I am not thinking about my phone number, even though I haven't

Talk of being unable to recall or bring a belief that not-*p* to mind, however, may not be entirely clear. Is it, for instance (1): the inability to recall *that not-p*, or is it (2): the inability to recall *that one believes that not-p*? If we look to the Beth case for assistance, we can see that what Beth eventually couldn't recall was *the proposition*: that her father loved her least. But it would also have been true that insofar as she forgot that, she also forgot that she believes that (since it doesn't seem to make much sense to say that she might have remembered that she believes that not-*p*, while not remembering that not-*p*). So forgetting a belief may involve both (1) and (2). It also seems that the explanation for why Beth forgot that her father loved her least is that she forgot *the reasons justifying* that belief, i.e. the memories of her father ignoring her for her brothers. If she had remembered those events, then presumably she would have 'remembered her belief': she would have remembered that her father loved her least. So saying that the belief that not-*p* is unconscious/inaccessible is saying that the subject has forgotten that not-*p*[52], and has done so because the reasons that made her believe that not-*p* in the first place won't come to mind.

Furthermore, philosophers who would maintain that the self-deceiver's unwelcome knowledge/belief is unconscious would not hold that it is unconscious in any stable, deep, Freudian sense, such that it might take psychotherapy for it to be retrieved (see Audi 1989: 249. McLaughlin 1988: 60-61, note 57). For the state of self-deception here is generally not regarded as a very stable one, and the buried belief is, so to speak, never far from the surface (i.e. it could have a tendency to come back to mind from time to time in response to certain

---

forgotten that number and could produce it on demand. This is not the use I am using here, and in fact this use is at odds with the historical one (e.g., Freud's) where an unconscious belief is an inaccessible belief.

[52] Saying 'he has forgotten that not-*p*' may have connotations of it being a purely unintentional matter, but these connotations should be ignored in this context since here the subject is responsible for it. Furthermore, this expression carries the implication that this belief of his is true. This is acceptable in this context since in the cases we are looking at, the belief that not-*p* is the unwelcome, warranted, and true belief, while the belief that *p* is welcome, unwarranted, and false.

cues, or under certain conditions). Accordingly, McLaughlin wanted to distance himself from the term 'unconscious', preferring 'inaccessible' which he hopes doesn't carry the same connotations of being deeply inaccessible. Audi, on the other hand, retains the term 'unconscious belief' but emphasises that it should be understood in a less extreme sense to the Freudian sense, as including beliefs that are inaccessible but not deeply so. It is important that we keep this construal in mind in the context of discussions of self-deception.

Some thinkers–Freud and Robert Audi (1982) for instance–would suggest a further difference between saying that a belief is unconscious, as opposed to eradicated, besides the implication that it is still possible for it to come to mind. They would claim that it is characteristic of unconscious knowledge/belief to be–though forgotten–still 'active', in the sense of causing, influencing, or manifesting in behaviour in certain subtle ways (or manifesting in 'neurotic symptoms' and dreams, in Freud's case). Though possible, whether this would be a universal feature of unconscious belief is debatable. For it could also happen that a belief of ours lies quietly 'dormant' in an inaccessible state, such that although the potential for it to become accessible remains, it does not manifest in or influence behaviour or thought. And we might also want to call a belief of this sort an unconscious belief. On Haight's analysis of unconscious knowledge, for instance, no mention is made of any necessary condition whereby that knowledge must be active, so that she accepts the idea of what we could call *dormant* unconscious knowledge.

There are many reasons why some knowledge might be unconscious in this sense, reasons that might have nothing to do with motivational factors. It might be that one has trouble recalling it because of natural memory loss due to it not being 'rehearsed' regularly in one's mind, for instance, or because it lacks personal significance. But regarding the forgotten knowledge/belief that is, in some accounts, associated with the self-deceiver, there is a special answer to the question of how that knowledge ends up being in that inaccessible

condition, namely, the self-deceiver '*make[s] himself* unconscious of what he knows and it is in this sense that he is responsible for the fact that he is not conscious of knowing this thing' (Hamlyn 1971: 56, emphasis added). Let's now look into how these theorists see this as being accomplished.

### *4. The Relation of the Notion of Unconscious Belief to Attentionalism.*

In Hamlyn's view, having contradictory beliefs is made possible by one of those beliefs being unconscious. And moreover, in self-deception, it must be that *the subject himself* is responsible for having made this belief unconscious, i.e. it must have been rendered unconscious as an outcome of his own actions (and in Hamlyn's view, the belief being unconscious may also be the *intended outcome* of those actions). Hamlyn's, and pretty much everyone else's answer to *how* the self-deceiver manages this is a typical statement of *the attentional strategy* which was looked at in the last chapter: '[he] can concentrate his attention on certain things to the exclusion of others, he can take a certain view of things to the extent that he becomes convinced of it and by it, he can make certain things explicit to himself rather than other things, and so on' (1971: 58). The idea is that by deliberately forcing one's attention off the evidence supporting a belief, repeatedly perhaps, one can render that belief inaccessible. Following tradition, we can call this the deliberate *repression* of the belief, insofar as it ends in the belief being inaccessible to consciousness.[53]

Is this the only way in which a person could render a belief of his unconscious? Probably not; one could probably imagine certain extraordinary ways in which one could do

[53] For Freud, at least in his later thought, repression was not something carried out deliberately by the subject. Rather he conceived it as an automatic mechanism triggered by anxiety. The notion of repression I am using here is different from this, it being a deliberate act, and not a mechanism. Also, Freud's use of 'repression' carries the implication that the repressed content is *deeply* inaccessible, such that it requires laborious therapy to make it accessible again. In this context, we should not take the term to carry this implication.

this, like if the subject had access to some appropriate technology. However, as I've said before, we are not interested in self-deception by 'special means' here, for such means are not available to the *ordinary self-deceiver*, which is the kind of individual we are seeking to understand. But if we confine ourselves to the kinds of ordinary means available to the average self-deceiver, it is not clear what other options there might be for accomplishing this besides the attentional strategy. At least no philosopher that I know of has come up with any other proposals in this regard, besides the idea that self-deceivers render their unwelcome belief unconscious by the control of attention. Now it could be that there is some other way self-deceivers can do this, and that we are lacking in imagination here, but it seems likely that if there was another way of doing it, it would have been mentioned. I am therefore inclined to say that the idea that self-deceivers have the capacity to render their own beliefs unconscious is *only as plausible as the doctrine of attentionalism* previously examined.

If the idea that self-deceivers have contradictory beliefs is dependent on the idea that they render one of those beliefs unconscious, and if the latter idea is dependent on the idea that they could do this by means of the attentional strategy, then the doctrine of contradictory beliefs will not look like a tenable position. For as we've seen, the empirical evidence suggests that the attentional strategy could not be effective, because people can't make themselves forget things by the control of attention (see chapter 6). What moves are then open to the traditionalist? The first option would be to try and break the second dependency by coming up with some other realistic way by which the self-deceiver could render a belief of his unconscious, different from the attentionalist strategy. As I've said though, there don't seem to have been any proposals in this regard, and I'm at a loss to think of one myself, so I won't be able to take this option much further. The second option would be to try and break the first dependency, by arguing that it needn't be the case that for one to have contradictory beliefs one of those beliefs has to be unconscious. This is a position that I believe some

would argue for, by adverting to the divisionism of Donald Davidson. For Davidson has been taken to hold that the self-deceiver has contradictory beliefs, but that neither of those beliefs is unconscious. In what remains of this chapter I will examine Davidson's divisionism and see if it does successfully break this dependency. I'll argue that contrary to what some of his interpreters say, Davidson does indicate some commitment to the idea that one of the beliefs has to be unconscious, and I will argue that in fact, this is the *only* way in which one could have contradictory beliefs.

## *5. Davidson's Divisionism.*

As I've said, a number of commentators have taken Davidson to have given no role to the supposition of unconscious beliefs or knowledge in his theory of self-deception (see Elster 1999: 427-428. Gozzano 1999: 143. Pears 1984: 83), though he did endorse the contradictory belief feature. Rather, this feature of the self-deceiver is supposed to be possible because one of the beliefs is 'kept apart', 'walled off' (2004/1986: 211), or 'partitioned' (2004/1982: 181) from the other. And a belief being partitioned from the other is, apparently, something different from it being unconscious while the other one is conscious. Davidson is usually spoken of as advancing a theory of 'mental division' here, also called 'divisionism'.

Ostensibly it may seem like divisionism offers an explanation of how it is possible to hold directly contradictory beliefs that is in competition with the explanation that supposes that one of the beliefs is unconscious or inaccessible. However, we should resist jumping to this conclusion. The reason is quite simple; the language of divisionism is, as Davidson himself admitted in his latest paper on the topic, *metaphorical*[54]. Insofar as it is metaphorical,

[54] The relevant sentence is 'I spoke of the mind as being *partitioned*, meaning no more than that a metaphorical wall separated the beliefs which, allowed into consciousness together, would destroy at least one' (2004/1997: 220).

it cannot offer us a *literal explanation* of how the contradictory belief condition is possible. To the question 'how can one believe that *p* and that not-*p*?' the answer, 'because the beliefs can be kept apart?' only invites the question 'what do you mean by "kept apart?"'. I will therefore take it to be important to identify what this metaphor is supposed to mean, since we can only pretend to know what we are talking about so long as the discussion is left at this metaphorical level. And it could turn out that when this interpretive work is done, we will find that Davidson's divisionism is not as different from other approaches as is sometimes thought.

Being metaphorical, we want to know what talk of the unwelcome belief being 'partitioned off' might be a metaphor *of*. The language of divisionism must be translated into more familiar notions that have a literal application if it is to be of any explanatory use, and if it is not to become a mere tool of equivocation. For this purpose we must take a closer look at Davidson's text for indications of how to translate this metaphor. Unfortunately, Davidson's remarks in this respect are scant, but they are not nonexistent, and there is material there for an interpretation to be constructed. However, his relevant comments on the status of the unwelcome belief, especially with respect to consciousness, are sufficiently ambiguous–if not inconsistent–to preclude conclusive conclusions from being drawn. I will, however, argue that the most charitable interpretation we can give to Davidson's view of the status of the unwelcome belief is that it is inaccessible, or unconscious in at least some weak sense, as it necessarily must be in my view.

It may first be pointed out that Davidson's divisionism is not to be interpreted along the lines of Pears' theory, which is also spoken of as divisionist (by Pears himself for instance), with the 'mental systems' that make up the person in that theory (the 'sub-system'

and the 'main system') counting as 'mental divisions'.[55] In Pears' version, at least one of the mental divisions, the so-called sub-system, is an agent and a subject of psychological attributions that is distinct from the person him/herself (it being something that is 'inside' the person). So when Pears attributes the intention to deceive, or the 'cautionary belief', to this 'mental division', he is attributing it to something that is distinct from or less than the person proper. Davidson never went this far with his notion of mental divisions, which for him didn't constitute different agents or psychological subjects. Rather, for Davidson a mental division (in the sense of one of the parts involved in the division) is, as will now be explained, simply a rationally related grouping of psychological attitudes, internally coherent among themselves, which are attributable to the person proper. So Davidson's theory is not open to the homuncularist objections that beset Pears' divisionism.

We can identify a preliminary explication of the notion of being mentally divided somewhat as follows. A person is mentally divided where that person has two propositional attitudes which are logically inconsistent with one another, or where one is irrational in light of the other (in an appropriately direct or conspicuous way). An example would be the belief that all things considered I should do *x* and the intention to do *y*, or the belief that *p* and the belief that this belief is unwarranted (he also calls this being in a state of 'internal irrationality'). So a mental division will consist of at least one attitude that is rationally inconsistent with an attitude in another division. Also, Davidson claims that because of the holism of the mental, each inconsistent attitude will be associated with or will presuppose a network of other attitudes that are rationally related to them, and many of these associated

[55] A quick note on the term 'division': This term is ambiguous, and perhaps even multiply ambiguous. It can refer to the fact that one thing has been divided into two or more parts (as in, 'a division was made in the group'). Or it can refer to any one of those parts involved (as with the military term 'a division'). And perhaps it can also refer to the physical gap or the dividing line between two divided things (in this sense a crevice may be a kind of division). I will usually be using the term 'mental division' in the second sense, and where I am using it in the first sense, I hope this will be evident from the context.

attitudes will be common to each division. For instance, the belief that man reached the moon can only be had by someone who has a number of other beliefs which 'support' and make intelligible that belief, including the belief that the Moon is in space, that it was reached in a spacecraft, that it's difficult to reach, etc., and the belief that man didn't reach the moon would presuppose and would be rationally connected with many of these same beliefs. Furthermore, the beliefs and attitudes in a mental division do not have any incoherencies among themselves. Divisions are, by definition, internally coherent.

It's worth noting that the point about the holism of the mental is not of great importance when thinking about Davidson's mental divisions. The reason is that there are such networks associated with *all* our beliefs and other psychological attitudes, but such networks don't constitute mental divisions in all these cases. My belief that Venus is the second planet from the Sun, and my desire for a curry, for instance, also presuppose or are rationally connected to quite different sets of propositional attitudes. But Davidson would not want to speak of such networks as constituting 'mental divisions'. The point that propositional attitudes presuppose 'networks' of other propositional attitudes is a point that applies to all propositional attitudes, rational or irrational, conscious or unconscious, and is not of significance for individuating mental divisions. What is important is the existence of the inconsistent attitudes, for this is what constitutes the *criterion*, or at least one essential criterion, for there being a mental division.

### *6. The Status of the Unwelcome Belief.*

Unfortunately, this does not make us any the wiser as to how the contradictory belief condition (or states of internal irrationality more generally) would be possible, nor does it tell us anything about the status of the inconsistent beliefs with respect to consciousness. For if

we limit ourselves to this understanding, to say that someone is mentally divided is just to restate, in a 'theatrical' or 'dramatic' way (Pears 1986: 85-86), that they have propositional attitudes which are inconsistent. And if the language of divisionism can only be explicated in this way, then as a number of commentators have remarked, the notion of a mental division cannot do any explanatory work for Davidson (Pears 1984: 84-85. 1991: 395-396. Gardner: 60-61. Borge 2003: 6), by which they mean, it can't explain how the internal irrationality is possible.[56] Going on what Davidson says in some places, one *could* be forgiven for thinking that he intends the language of divisionism to be understood only in this way, (e.g. 'the breakdown of reason relations *defines* the boundary of a subdivision' (2004/1982: 185, my emphasis)). On this understanding, mental division would not be–as Davidson sometimes suggests–something further that is *postulated* or *hypothesized* (2004/1985: 198. 2004/1986: 211) to *explain* the possibility of contradictory beliefs, for the fact that there are contradictory beliefs would simply constitute the fact that the person is mentally divided.

The inadequacy of this explication of divisionism should make us return to the question of whether Davidson thinks that one of two contradictory beliefs which the subject holds simultaneously is unconscious, since if Davidson was suggesting that, this *would* be truly explanatory of how this 'internally irrational' state is possible. I suggest that the textual evidence that Davidson did reject the idea that one of the beliefs must be unconscious is ambiguous at best. Certainly, there are remarks that strongly suggest such a rejection. Most importantly, Davidson explicitly says that the unwelcome evidence is *not forgotten* (2004/1986: 209-210). And if that information is not forgotten, then it should be accessible to influence the subject's beliefs.

[56] This is why Pears felt the need to make the further, radical step of conceptualizing mental divisions as separate agents and psychological subjects. By doing so, Pears regarded his divisionism as an improvement over Davidson's by being truly explanatory of how internal irrationality is possible and as being no longer open to the objection of triviality, though Pears approach, as we saw, sufferers from problems of its own.

But there are reasons to be wary of this interpretation. Firstly, it was in his earliest paper on irrationality, *Paradoxes of Irrationality*, that Davidson spoke about not having a need for the unconscious in his analysis, but there he was referring to weakness of the will or *akrasia*, and not self-deception (2004/1982: 185-186). Self-deception was not discussed in that paper (only mentioned), though wishful thinking was. Davidson believes that the *akratic* individual does some action *A* while also thinking that *A* is not the best thing for her to do, all things considered. He then insists that in central cases, her belief that *A* is not the best thing for her to do is not unconscious; she is aware that *A* is not best for her to do, but she does it anyway. But from the fact that Davidson saw no role in this form of irrationality for 'the unconscious', we can't conclude that he saw no role for it in self-deception. And if Davidson would have denied that the unwelcome belief is unconscious, the point might only be that it is not unconscious in the *Freudian* sense of being *deeply* inaccessible to consciousness. This denial would be consistent with maintaining that it is unconscious in a weaker sense of being inaccessible to some extent or other.

It is better to look to those papers where he discusses self-deception directly, namely, *Deception and Division* and *Who is Fooled*? for Davidson's considered view on whether the unwelcome belief is inaccessible or not. There, another current of thought can be identified. On the few occasions where he hints at a clarification of what it means for one belief to be 'kept apart' from another, he spoke in terms of consciousness. The most important sentence is: 'I spoke of the mind as being *partitioned*, meaning no more than that a metaphorical wall separated the beliefs which, allowed into consciousness together, would destroy at least one' (2004/1997: 220. Also see Davidson 1999: 444). Furthermore, in a more metaphorical tone he said that 'the agent cannot survey the whole without erasing the boundaries' (2004/1986: 211). By 'the whole' Davidson here seems to mean the two beliefs, or all the evidentially relevant considerations, and 'survey' here looks like a synonym for 'be conscious of'.

In these sentences Davidson's attention is turned towards the possible occasion where the subject who believes that *p* and that not-*p* actually *consciously thinks about the p-issue*. For Brian McLaughlin, this is where the real puzzle about the contradictory belief feature lies. For we want to know what implications this idea would have for an occasion when the topic of whether *p* was to come to the subject's mind, since it is in this situation that the opposed beliefs would be expected to 'clash' (as he puts it) with one another. The problem follows from some plausible assumptions about what one could legitimately expect a subject to think, given the fact that he believes that *p*, were he to consciously consider the issue of *p*, assumptions that have been endorsed by McLaughlin (1988), and also Kent Bach (1981). They are as follows. Take *p* to be 'Santa exists'. As we all know, Jones can believe this though not be consciously thinking that Santa exists, or of anything about Santa. Nevertheless, given that Jones does believes that Santa exists, then if Jones were to consciously think of the issue of Santa's existence, we would expect him to consciously think that Santa exists (I will drop the 'consciously' from now on). That is:

If *S* believes that *p*, then were *S* to think of whether *p*, he will tend to think that *p*.[57]

Or as Bach put it, 'to believe that *p* is to know, should the thought of *p* occur, what to think about it without having to deliberate' (1981: 355). The 'will *tend* to think that *p*' here should mean something like this: If *S* believes that *p*, *S will* think that *p* if he thinks whether *p*, *unless something prevents him from doing to*. That is, the fact that *S* doesn't think that *p* when he

[57] McLaughlin and Bach phrase this differently. They say that given that Jones believes that *p*, if Jones thinks *of p*, he will tend to think *that p*. But the expression 'think of *p*', strictly speaking, is grammatically malformed, which we can immediately see once we substitute a specific proposition for *p* (e.g. 'Jones thinks of Santa exists'). It seems to me, however, that we should understand 'thinks of *p*' as a contraction for 'thinks of whether *p*' for it seems both grammatical and true to say that if Jones believes that *p*, and thinks of whether *p*, then he will tend to think that *p*.

thinks whether *p* is normally good reason to say that *S doesn't* believe that *p*. So if *S* does in fact believe that *p*, and yet doesn't think that *p* when he thinks whether *p*, there must be some special explanation for why that happens. Something must have gone wrong; something must be blocking or inhibiting the natural disposition for the belief to manifest itself in an occurrent judgment that *p* in these naturally activating circumstances.

For McLaughlin the above conditional statement presents a *prima facie* difficulty for the notion of contradictory beliefs, because the question now arises of what would happen if the person with these beliefs were to consciously think of whether *p*. He considers a case where *x* believes he is overweight and believes he is not overweight.

> Suppose that *x* were to think of his being overweight. Then...would he think that he is overweight, or that he is not overweight, or would he just not know what to think? What he *cannot* do is think that he is overweight and that he is not overweight. So, it would seem that if the topic of his being overweight came to *x*'s mind, *x* would either lose one of the two beliefs or lose both (1988: 48).

I would imagine that Davidson would concur with McLaughlin here. In the two Davidson quotations we looked at earlier, he suggests that both beliefs can't be 'in consciousness together'. If what it characteristically is for a belief that *p* to 'be in consciousness' is for the believer to be thinking that *p*, then the impossibility of the two beliefs being in consciousness amounts to the impossibility of one thinking that *p* and that not-*p* at the same time, which is the impossibility McLaughlin speaks about. So the picture here is as follows: if the self-deceiver, *S*, believes the welcome proposition *p* and the unwelcome one, not-*p*, then if he were to consider the issue of whether *p*, then he would be inclined to think that *p*, and not think that not-*p*, a thought that has been suppressed.

By 'inclined' here I intend to leave open the possibility that on *some* occasions when *S* is thinking of the *p*-issue, he may think that not-*p*, and not think that *p*. The self-deceiver may experience moments of lucidity towards the truth within a period where he would be generally classed as self-deceived. This may be consistent with saying that he was self-deceived over that period, so long as this tendency towards thinking that not-*p* did not prove dominant over the tendency to think that *p* when considering whether *p*. Thus saying that *S* is self-deceived is like saying that *S* likes his job, which is not inconsistent with the fact that *S* may get frustrated or bored with his job from time to time, so long as those feelings don't dominate over his more positive ones.

McLaughlin would call the belief that not-*p* here an 'inaccessible' belief. For McLaughlin, a belief that *p* of *S*'s is inaccessible if and only if *S* believes that *p*, but were *S* to think of the issue of whether *p* he would not think that *p*. So if *S* believes that *p* and that not-*p*, if he thinks of whether *p*, *at least one of these beliefs will be revealed to be inaccessible, since S cannot consciously think that p and that not-p simultaneously*. This is why it is true that for one to have contradictory beliefs, one of them *must* be unconscious/inaccessible, if these terms are defined as McLaughlin has done. I would suggest, furthermore, that Davidson is committed to such a picture, since Davidson's remarks suggest that one of the conflicting beliefs must not arise in consciousness if the *p*-issue arises or is contemplated, and that would simply entail that it is inaccessible under McLaughlin's reasonable use of that term.

It looks like we are now in a better position to attach a more substantive meaning to the language of divisionism. For Davidson, what it is for two beliefs to be 'kept apart' is for one not to come into consciousness when the other is in consciousness, due to it being suppressed. The unwelcome belief doesn't come into consciousness because the thoughts or memories of the unwelcome considerations which caused the subject to have the belief (and which justify him in having it), have been suppressed and won't come to mind when he is

contemplating the *p*-issue. The so-called 'boundary' or 'mental partition' between 'mental divisions' is then a metaphor for this fact.

## *7. A Possible Objection.*

Yet one may object that this does not imply that the knowledge of the unwelcome evidence, and the belief that not-*p*, is inaccessible. The objector here may refer to the distinction between some knowledge, or a belief, not being access*ed*, and not being access*ible*. The former just means that it isn't in consciousness, the latter implies something like, that it *won't* come to consciousness, or that the subject is *unable* to bring it to mind, or if we are talking about memories (e.g. of the evidence), that it is *forgotten* or that the subject *can't remember* it.[58] Davidson implies that when the self-deceiver thinks of the *p*-issue (or at least much of the time when he does), he thinks that *p*, and thoughts of the opposing evidence, which he did know about, don't come to mind (where if they did come to mind, he would not think that *p*, but that not-*p*). But mightn't this only mean that the knowledge of the evidence, and the belief that *p*, is not being accessed as opposed to being inaccessible? A minute ago, for instance, your weren't thinking about your phone number, but you weren't unable to bring it to mind, or you hadn't forgotten it, for all that. It wasn't *being accessed*, though that doesn't imply it was *inaccessible*. However, what I want to say is that if thoughts of the unwelcome evidence did not come to consciousness *at any point when the subject was thinking of the p-issue*, this *does* entail that the subject forgot about this evidence (and hence that his

[58] I do not think that 'inaccessible' means 'forgotten'. If we can talk about inaccessible (or unconscious) beliefs or desires, then it can't mean forgotten, because it's not right to speak of having forgotten a belief or desire (or rather, when we say such a thing, we mean something different from what's intended here). But when we speak of *memories* being inaccessible, or often of *knowledge* being inaccessible (like knowledge of facts), their being inaccessible means, or at least implies, that these memories or this knowledge is forgotten. Saying a memory is forgotten is equivalent to saying that it won't come to mind or that the subject can't bring it to mind.

knowledge of the evidence was inaccessible at that point). Or at least it does under an understanding of the notion of forgetting which I would endorse as being plausible, according to which information has been forgotten if it doesn't come to mind in circumstances when that information is *relevant to the situation*. I will elaborate on this point now.

Let us attempt to establish the conditions under which some knowledge of ours, or a memory, would customarily be said to have been forgotten. Consider Burke, who learned about a shortcut for getting from A to B. Now from the fact that the matter of the shortcut from A to B isn't 'in his consciousness', or in layman's terms, isn't on his mind or in his thoughts, we cannot conclude that he has forgotten about the shortcut. However, if the thought of the shortcut doesn't enter his mind *at a time when that information is relevant*, then we have grounds for saying that he has forgotten it. For instance, if Burke is travelling from A to B (or from B to A) then that is a time when that information would be relevant, and if it didn't enter his mind then (and he ended up taking the long route because of it), it would be natural to say that he forgot all about the shortcut. Another occasion when this information would be relevant would be if someone were to ask Burke, 'do you know of any shortcuts from A to B?', and if the information didn't come to mind here, we would have grounds for saying he forgot it (note that though Burke might not remember *the information*, he might still remember *that he knows* the information, and might try, in vain, to recall it). The crucial notion here is *circumstances when that information is relevant* (or when the remembering is 'to be expected', as Haight put it). To see how crucial this notion is, consider that Burke might not have thought about the shortcut for 30 years, and that fact could provide no grounds whatsoever for asserting that it was forgotten. This would be so if he moved country for 30 years, since in another country that information might never have been relevant. We could not infer from his not thinking about the shortcut for those 30 years that he forgot that information, because it could still have been true that had one asked him 'do you know a

shortcut from A to B?' he would have immediately been able to tell him about the shortcut. And that would be grounds for saying that he had never forgotten it. But as soon as Burke is in a situation where that information is relevant, if that information doesn't come to his mind, we have grounds for saying he has forgotten that information.

Let us apply these thoughts to the case of the internally irrational self-deceiver. Davidson says that the self-deceiver cannot have both beliefs in consciousness together. If the welcome belief that *p* is in consciousness (that is, if he is consciously thinking that *p*), then he can't think of the evidence against this, which would make him think that not-*p*. However, an occasion when the subject has the belief that *p* in consciousness is an occasion when the subject is thinking about the *p*-issue. And the unwelcome evidence is *relevant* to the *p*-issue, in the sense of being *evidentially* relevant, and indeed, by hypothesis, evidentially *crucial* to it. So if the subject thinks of the *p*-issue, but *C*- does not come to his mind, then it follows that he has forgotten *C*-, at least at that time. Again, the reason is that if in some situation certain previously learned information does not arise in consciousness when it would be relevant to that situation, those are the circumstances in which we count this information as having been forgotten.

So if Davidson's language of divisionism is to be explicated in a way that makes it capable of explaining how it is possible to believe that *p* and that not *p*, we must interpret him as holding that for the self-deceiver, the unwelcome belief is unconscious or inaccessible to consciousness to some degree, at least some of the time. So a belief that has been 'walled off' is one that has been rendered inaccessible (where this notion can be understood in terms of the inability to recall or bring to mind the evidence supporting that belief, because it has been suppressed). This interpretation follows from some remarks of his, coupled with a particular understanding of what it means to forget something (and this, let it be noted, is not to attribute to Davidson the view that the unwelcome belief is unconscious in any deep,

Freudian sense). To my knowledge, nobody has come forward with any other clear, workable suggestions for interpreting divisionism, and as Annette Barnes has said who endorses a similar interpretation, it is difficult to see how else the notion of 'partitioned off' be understood besides in terms of the notion of forgetting (1997: 30).

## *8. Concluding Remarks.*

In summary, I have tried to argue that the only viable answer to the question of how the contradictory belief condition is possible involves holding that one of the beliefs is inaccessible to consciousness. For the traditionalist, the subject herself renders this belief inaccessible as part of her effort to deceive herself, and the strategy she uses for this is supposed to be the attentional strategy. However, in chapter 6, we saw that the attentional strategy would not be a viable option to the self-deceiver. Ordinary people just do not have the power to render their own beliefs inaccessible. Therefore, it seems that ordinary people couldn't intentionally get themselves into a state where they believe that *p* and that not-*p*, and so this is not a feature of ordinary self-deception.

In the next and final chapter, I wish to make some final adjustments to my account so that it can meet an important desideratum of a good theory of self-deception: that of being able to explain the 'tension' of self-deception.

# Chapter 8: The Tension Inherent in Self-Deception

## *1. Introduction.*

In this chapter, I wish to turn from the critique of the last few chapters to developing further the positive account of self-deception being promoted here, since there are a few loose ends that need to be tied up with it. This development is being made in response to a criticism frequently made towards Mele's account, or any account that represents the self-deceiver as straightforwardly believing something falsely/unwarrantedly and being ignorant of the truth, and I anticipate that this same objection would be made to my own view, since I have represented the self-deceiver as having an unwarranted belief that *p*, and as not believing the truth that not-*p*.

The criticism is that accounts of this sort fail to capture or explain the 'tension' inherent in self-deception (Funkhouser 2005: 299, Nelkin 2002: 391, Audi 1997: 104, Bach 1997: 105, da Costa & French 1990: 182-183). Some of these thinkers then go on to say that this tension can be best explained by supposing that self-deceivers have contradictory beliefs, whereas others hold that it can be best explained by supposing that self-deceivers do believe the unwelcome truth and don't believe the unwarranted proposition which they are disposed to avow at all. However, I wish to avoid attributing contradictory beliefs to the self-deceiver, and I also want to accommodate the intuition that self-deception involves having an unwarranted attitude of some sort.

As I will point out, however, there are important differences in the ways in which this tension idea has been explicated and illustrated by philosophers, in light of which it is doubtful that there is a single explanandum we should have in view here which would require a single kind of explanation. Accordingly, two main ways of understanding this tension idea will be distinguished in this chapter, and it will be argued that on one such understanding,

deflationism is well positioned to explain the phenomenon. This, however, will require some modifications to the account given. These modifications will be informed by some observations made on experimental work done on the biasing influence of desire on belief–empirical work from which Mele's deflationism has received much of its inspiration–and will involve supposing that the self-deceiver has an unwarranted degree of conviction in a proposition rather than an unwarranted belief in it.

However, there is another way that the idea of tension has been understood by philosophers, a way which, as I will argue, is incompatible with the idea that self-deceivers have an unwarranted attitude of any sort. This understanding can't be reconciled with the account being recommended here. But the phenomenon that philosophers have in mind here is an entirely different phenomenon to self-deception, or so I will argue.

## *2. The idea of Tension.*

So what is meant by 'tension' here? To answer this question we can only look to the ways this idea has been explicated by the philosophers in the literature who accuse Mele's deflationism of being unable to explain it. Unfortunately, remarks on this tend to be quite scant and sometimes vague, and as I've said, the ways the idea has been cashed out by different philosophers are not always consistent with one another. Nevertheless, I will try to group together some remarks which seem to indicate a single kind of phenomenon which we can then treat as the relevant explanandum.

Eric Funkhouser sees the tension of self-deception as including both 'cognitive and behavioural tension' (2005: 296). 'Behavioural tension' at the most general level refers to the alleged fact that the self-deceiver displays some behaviour that seems more consistent with believing that *p* and other behaviour that seems more consistent with believing that not-*p*.

Besides this, self-deceivers supposedly also experience *mental* or *cognitive* 'conflict', 'tension', or 'discomfort' (Funkhouser 2005: 299. Graham 1986: 226. Losonsky 1997: 122). Graham elaborates this as the experience of being afflicted with 'doubts, qualms, suspicions, misgivings, and the like' concerning the belief we are self-deceived in holding (1986: 226), or in Losonsky's words, 'recurring and nagging doubt' (1997: 122, also see Funkhouser 2005: 299). Noordhof similarly speaks of an essential 'instability' present in the self-deceived state (2009). I would assume that these philosophers think that such mental tension is the experiential accompaniment for those cases in which behavioural tension is present or liable to occur. They are two sides of the same coin.

Assuming that this phenomenon is an important feature of the self-deceived condition, what kinds of attitudes should we ascribe to the self-deceiver to explain it? Many thinkers who stress the idea that self-deception involves such tension have seen it as giving us grounds for attributing contradictory beliefs to the self-deceiver. Graham, for instance, says that '[t]he supposition that self-deception requires [believing *p* and not-*p*] can help to account for the discomfort of self-deceivers... With [believing *p* and not-*p*], discomfort can be expected' (1986: 228). A number of others have also used the supposition of contradictory beliefs to explain this behavioural and mental tension (see da Costa & French 1990: 183. Demos 1960: 591-592. Dion Scott-Kakures 1996: 48-49. McLaughlin 1988: 51. Steffen 1986: 132-133). This would be incompatible with deflationism, however, which we partly defined through its denial that self-deception involves having contradictory beliefs. Alternatively, others have suggested that in light of such tension, there may be no determinate answer to the question of what the self-deceiver believes (e.g. Funkhouser 2009, Hamilton 2000: 25).

Mele's characterization of the self-deceiver, on the other hand, is roughly one of a person who desires that *p*, and who falsely believes that *p* against good evidence to the contrary, because that desire has caused him to 'treat the data' in a biased way. Thus Mele

straightforwardly represents the self-deceiver as believing the false, unwarranted proposition, and not believing the truth, and this has led to the perception among Mele's critics that the mental state he associates with self-deception is tension free.[59] Some theorists prefer to reserve the term 'delusion' for this believing without tension against the evidence, presumably because the stability and surefootedness of this apparently tension-free belief would seem to indicate an insensitivity to reason that may be more the mark of pathology (see Funkhouser 2005, Graham 1986, da Costa & French 1990, Demos 1960).

I take it that we have identified in the above remarks, under the heading of "tension", a more or less definite phenomenon, which involves on the behavioural side of things: being inclined to act in some ways that seem more consistent with believing that *p* and in others that seem more consistent with believing that not-*p*, and then, on the connected mental side of things: being afflicted by 'doubts, qualms, suspicions, misgivings, and the like', or 'recurring and nagging doubt'. So we will take the objection to be that tension, in this sense, is an important feature of the self-deceived condition, and that a deflationary theory of the sort that Mele recommends, which rejects the contradictory belief feature, cannot account for it. My strategy for meeting this objection will not be to try to argue against the idea that such tension is characteristic of self-deception, but to affirm that it is, and to modify the deflationist theory in such a way as to account for it. This modification will derive from the consideration, in the next section, of some relevant empirical work. Later, I will describe another way in which the tension idea has been understood, and my strategy here will be to argue that tension in this sense is not characteristic of self-deception, and hence is not something we are obliged to explain.

[59] Mele does say that his ideas on self-deception could be formulated in terms of 'degree of belief/confidence' (2001: 10). However, he doesn't exploit this possibility for dealing with the tension issue, though he perhaps suggests this approach when he suggests one way of accounting for behavioral tension cases by saying the self-deceiver may believe that *p* while believing there's a significant chance that not-*p* (1997a 96).

3. *Unwarranted Degrees of Conviction.*

Recall now the study by Kunda (1987) discussed earlier in which subjects had to read a scientific article warning of the dangers of caffeine for women. In this experiment, the stakeholders (female high caffeine consumers) and the non-stakeholders (female low consumers and male high and low consumers) came to form different attitudes concerning the veracity of the article. Stakeholders were more skeptical of the article, and of the allegation that there is a causal connection between caffeine and ill-health in women, than non-stakeholders. However, I also pointed out that the difference in attitudes between the two groups was not described (and indeed, was not describable) in terms of outright belief, or in terms of what we might call the *belief that p/not-p schema*. Rather, Kunda probed the attitudes of these subjects by asking them to indicate on a questionnaire *how convinced* they were of the proposition in question, where they had to put a mark on a scale. At the time, I put this matter aside, and for the sake of ease of exposition I adopted the usual practice of discussion self-deception in terms of outright belief rather than in terms of a notion allowing for gradation, such as degree of conviction in a proposition. However, now I would like to return to this issue, since as I will argue, it is of much relevance to understanding the tension of self-deception.

Mele appeared to present Kunda's study as a demonstration of self-deception, and so may have been presuming that the stakeholders in it satisfy his criteria supposed to be sufficient for self-deception. However, it is doubtful that they do so when we consider how Mele framed his conditions in terms of outright belief. Consider, for instance, Mele's Impartial Observer Test for self-deception. For Mele, where *S* believes that *p* and desires that *p*, we have good evidence that this belief has been biased by her desire that *p* when she passes the following test:

> ...given that *S* acquires a belief that *p* and *D* is the collection of relevant data readily available to *S* during the process of belief-acquisition, if *D* were made readily available to *S*'s impartial cognitive peers and they were to engage in at least as much reflection on the issue as *S* does and at least a moderate amount of reflection, those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true (Mele 2007: 167).[60]

In fact, the logic behind Mele's impartial-observer test is very similar to the logic behind Kunda's experiment. There, non-stakeholders approximated to Mele's impartial cognitive peers (ICPs). They were 'impartial' in that they had no personal stake in the issue (or at least less of a personal stake than the stakeholders). They were 'cognitive peers' in that they judged on the basis of the same body of data as stakeholders (i.e. the article), and also in that for at least some of them, we can presume there were no significant differences in relevant background beliefs (and cognitive ability) compared to stakeholders.[61] Kunda ensured this by having the alleged ill effects apply only to women, who presumably had the same prior beliefs about caffeine as the male heavy caffeine consumers in the group. The judgments of these males also differed from the stakeholders, and to the same degree as the other non-stakeholders.

---

[60] Besides this, the belief that *p* would have to be false, and the desire would have to have biased the belief in a 'non-deviant' way for a case to meet all his sufficient conditions for self-deception.

[61] Note that to qualify as ICPs, the stakeholders would also have to put a reasonable amount of effort into assessing the issue. The idea here is presumably that we should not put much stock into the judgments of people who assess an issue in a cursory way, as people for whom the issue does not matter could be inclined to do. Rather, an ideal judge would be one who is impartial, and yet motivated to come to an accurate judgment on the issue. Though in this experiment Kunda did not seem to make any special efforts to ensure that non-stakeholders were so motivated, the experimental context might have supplied this somewhat. Furthermore, other studies that have made such efforts have resulted in similar deviations between the judgments of stakeholders relative to non-stakeholders to what's witnessed here (Lundgren & Prislin 1998).

However, closer examination of the study reveals that stakeholders fall short of passing the test in an important respect. Kunda had the subjects indicate on a 6-point scale how convinced they were of the purported link, where '1' meant 'not at all convinced' and '6' meant 'extremely convinced'. The difference between the attitudes of stakeholders and non-stakeholders, though statistically significant, was not gaping. Stakeholders' level of conviction averaged at about 3 and non-stakeholders' level of conviction averaged at the 3.5 mark. In a later replication of this study, on a 9-point scale stakeholders averaged at 5.6 and non-stakeholders at 6.72 (Liberman & Chaiken 1992: 674). It thus appears that the discrepancy between the averaged judgments of stakeholders and non-stakeholders was subtle, and is not one where the former believed that *p* while the latter believed not-*p*. Furthermore, these results are similar to those found in other studies on motivationally biased belief (e.g. Wyer and Frey 1983).

Does this mean that the stakeholders do not pass Mele's test for self-deception? Perhaps so, though we would really have to look at them on an individual basis here in line with the test (Kunda's article, however, only supplies data on them as a group). However, this only highlights the point that Mele's test may be too stringent. For consider a case where the difference in attitudes between a stakeholder, *S*, and her ICPs is described in terms of their degree of conviction. Say that on a scale from 0-10 her ICPs would on average mark 9 (which might be a rough numerical translation of their being 'fairly sure' that not-*p*) while she would mark 5 (meaning, perhaps, that she's 'rather uncertain' that not-*p*). Though this is not a case of her believing that *p* while her ICPs mostly believe that not-*p*, her attitude towards the proposition would nevertheless have displayed a noteworthy, desire-driven deviation from what's warranted by the evidence, and this, I propose, should still attract a charge of self-

deception.[62] As I've said before, if we are prepared to accept that believing that *p* when the evidence warrants the belief that not-*p*, where this is caused by a desire in the appropriate way, is sufficient for being self-deceived, then we should be prepared to accept that having an *unwarranted degree of confidence* in a proposition, where that is caused by a desire in the same appropriate way, should also qualify one as self-deceived. For this would be just a less extreme variety of essentially the same phenomenon.

### *4. The Notion of Degrees of Conviction.*

Let me say a little about this notion of degrees of conviction which I see as being in play here. Again, the psychological studies looked at above suggested this notion to us through using questionnaires on which subjects had to indicate how convinced they were of the relevant proposition. Moreover, this seems to have been a meaningful question to have asked them. Subjects apparently understood and followed these instructions, and reliable results ensued where stakeholders were found to have marked on average lower down the scale then non-stakeholders, as was expected. These facts suggest that on some occasions there is a more fine-grained way of capturing the attitudes of subjects towards propositions than can be accomplished with the coarser conceptual apparatus of outright belief.

The notion of degrees of conviction/confidence (often called 'degrees of belief'[63]) is one that, according to Eriksson and Hájek (2007), has resisted philosophical analysis. That in

[62] In fairness to Mele, nothing he says is incompatible with this point, since he was only trying to establish sufficient conditions for self-deception when he proposed this test.

[63] This choice of expression has some disadvantages I believe, and is one I wish to avoid. Theorists who employ this terminology often represent the range of degrees of belief on scales ranging from 0 to 1, where 0 means that one is certain that not-*p*, and where 1 means one is certain that *p*. However, many also believe that talk of believing that *p simpliciter* can be understood as having a sufficiently high degree of belief. Consequently, one can fail to believe that *p* because one has too low a degree of belief that *p*. But this is odd; one would think that

itself is no reason to reject the notion however (many patently legitimate notions defy such attempts), and like those authors, I will take it as a datum that there are such things, represented as they are in such everyday locutions as when we claim to be or feel fully convinced or certain, very convinced, fairly convinced, not very convinced, not at all convinced, etc., that *p* (where we sometimes use 'confident' or 'sure' instead of 'convinced').

It is also possible to try to represent our degree of conviction numerically, which allows for much finer discriminations than would be possible with the more colloquial categories of 'fairly', 'very' etc. It is, however, important to recognize that often such scope for fine differentiation will be superfluous, and there will be indeterminacy concerning the question of what the appropriate numerical value should be to express our degree of conviction. For instance, though I might claim to feel fairly confident that a certain team will win a football game, I might not be able to decide on whether my confidence would be best represented by a .7 or an .8 on a scale going from 0 to 1, much less decide it to the second decimal place, and it might be that no 'indirect' test could establish which would be more appropriate, for there simply may be no fact of the matter as to whether .7 or .8 would capture my degree of conviction more correctly. However, this would not mean that I have no degree of conviction here at all, for I might not hesitate in saying that a .8 captures it better than a .2. It would only mean that degrees of conviction can't always be finely and precisely differentiated in that way. It is perhaps because of this that psychologists typically use scales of less than 10 points when measuring attitudes.

Besides relying on avowals as a measure of a subject's degree of conviction, we should also note that how convinced one is of a proposition will have implications for how much one is willing to risk on the assumption that it's true, and consequently, one's behaviour in circumstances where there are things to be gained or lost from acting on that

---

having any degree of belief towards *p* presupposes that one believes that *p*, just as feeling angry towards A to some degree presupposes that one feels angry towards A, even if only a little.

assumption is another important measure of one's degree of conviction (I will discuss the connection between belief and risk taking some more later).

There are many important philosophical questions about degrees of conviction. For instance, there is the important question of how degrees of conviction talk relates to belief talk. Perhaps the most common view here is that the concept of believing that *p* is that of having a high degree of confidence in *p*, with some arguing that the concept of belief is vague on what counts as high (Foley 2009, Hunter 1996). There are also questions on how degree of conviction talk relates to talk of belief in how likely the proposition is, and on the place this notion has in the explanation of action. Luckily, we should be able to get by here without having the answers to all these questions. I hope, therefore, to have said enough to give some degree of clarification to the notion of degrees of conviction/confidence for current purposes. Let us now turn to the issue of what use this notion can be in helping us to explain the tension inherent in self-deception, and of how it can be integrated into the deflationary account.

## *5. The Payoff with Changing from Talk of Belief, to Talk of Degrees of Conviction.*

Rethinking self-deception in terms of degrees of conviction would bring a number of advantages. Firstly, this more discriminating descriptive vocabulary allows us to talk about cases where a subject's attitude has not deviated from that of her ICPs such that the former believes that *p* while the latter believes not-*p*, but in a way that is substantial nevertheless, and that intuitively permits a charge of self-deception. It may be important for us to be able to do this, because if the above psychology experiments into motivationally biased belief are anything to go by–experiments which are fairly representative in this respect–then deviations in confidence level between self-deceiver and ICP of a limited magnitude may be the norm when directional desire influences belief. Imaginary cases which turn up in the philosophical

literature where the self-deceiver believes outright that *p* while his ICPs believe that not-*p*, while heuristically useful, promote more exaggerated ideas about our powers for self-deception than would be suggested by these results.

Secondly, rethinking self-deception in these terms allows us to account for the tenacious conviction that self-deception involves tension, and to do so without appropriating controversial notions such as that of contradictory beliefs. Again, this tension is described by some as the condition of being afflicted with 'doubts, qualms, suspicions, misgivings, and the like', 'recurring and nagging doubt', and 'instability', and as being manifested in ambivalent behaviour. If one were trying to explain this while wedded to the *p*/not-*p* schema, then the hypothesis of contradictory beliefs could seem like one of the few options available for making sense of this, since neither the supposition that the self-deceiver believes exclusively that *p* or exclusively that not-*p* would seem capable of explaining this tension. (It's not the only option though. As I've said, some have from this tension supposed that there's no determinate answer to the question of what the subject believes). However, a more natural explanation is indicated by the empirical studies. These showed deviations in confidence level between self-deceivers and ICPs that are noteworthy but also constrained by the real force of the evidence, suggesting that ordinary people do not have the kinds of powers for self-deception that would lead them, through processes of biased reasoning etc., to form stable, settled beliefs that completely contradict that of their ICPs. The inclusion of scales that probe confidence levels in these studies then suggests another reason why self-deception involves tension. Self-deceivers simply may not manage to fully convince themselves of what they want to be true. As Kunda remarks, their concerted efforts to construct justifications for their preferred positions are constrained by considerations of plausibility (1990: 482-3).

In a state where one's confidence level ranges between wholehearted belief and disbelief, it would be natural to expect the aforementioned 'tensions' to appear. For instance,

on the behavioural side, it is natural for people to make allowances for the possibility that *p*, and also to make allowances for the possibility that not-*p*, when they are uncertain as to whether *p*. For instance, to work with a type of case often mentioned in the literature, if one felt uncertain as to whether one would soon succumb to one's illness (and one might be self-deceived in so feeling, if the relevant evidence suggested that this was an inevitability), then it would be natural for one to make allowances for what one sees as two live possibilities. One might, for instance, draw up a detailed will on the chance that not-*p*, though one might also book a holiday for the summer on the chance that *p*. Such 'ambiguous' behaviour may be in fact rational for this person in light of what she thinks is likely, though she may nevertheless be *irr*ational, and self-deceived, in how likely she sees her chances of recovery as being. So if the "behavioural tension" of self-deception is to be understood in terms of the self-deceiver acting at times as if on the assumption that *p*, and at times also as if on the assumption that not-*p*, it can be accounted for in terms of unwarranted degrees of conviction in a proposition without need for the supposition of contradictory beliefs (though as we will see, some philosophers may want to understand the behavioural tension idea differently).

With regards to one's phenomenology, we could here expect "mental tension" to arise too, if this be understood along the lines previously mentioned ("doubts, qualms, suspicions, misgivings, and the like", "recurring and nagging doubt" etc.). It is true that uncertainty in itself does not cause mental tension; many propositions we are uncertain about give rise to no such experience. But the idea here is that uncertainty *combined with* the fact that one has a *stake* in the issue makes for the difference between merely having doubts about something, and *feeling plagued* or *nagged* by those doubts. The self-deceiver struggles to justify and find reasons for her favoured position through her biased thinking, but she does not entirely succeed in countering the unwelcome evidence to her own satisfaction. Because of the stake she has in whether *p*, her doubts as to whether *p* are a source of *worry* for her, which they

would not be for someone with those same doubts but without such a stake. They plague or nag her, but not a non-stakeholder in the same doxastic position.

Note that by mental tension here, we are not referring to something unique to the self-deceived state. It is common to feel tense in situations where we are uncertain about something that we have a stake in. Think of the state of tension a man may be in when watching a horse-race when he has money on one of the horses, for instance. This tension is a natural product of those two factors of uncertainty and concern being combined, though a person need not necessarily be self-deceived in situations of this sort. The behavioural phenomenon discussed is likewise not something special to self-deception.

I will now offer an alternative deflationary analysis of self-deception, which like Mele's analysis is only meant to give jointly sufficient conditions, and that is supposed to characterize ordinary, straight self-deception. *S* is self-deceived when:

1) *S* desires that *p*, and encounters data (evidence or considerations) that challenges, to some extent, the assumption that *p*.
2) His desire that *p* leads *S* to treat this data (reason about it and evaluate it), and to search for further data, in a biased way.
3) This biased treatment (non-deviantly) causes *S* to have a degree of conviction in the proposition that *p* greater than what is warranted (by the data which *S* possesses, and which was easily available to *S* to possess), to a noteworthy extent.

Let me make a few clarifying remarks on this set of conditions. The idea of 'encountering data' in condition (1) should be taken to imply that *S* comes across the evidence and *appreciates* that it is something that, on the face of it, poses a challenge or threat to the assumption that *p*, for this is necessary to motivate the attempts to deal with it (by, for

instance, trying to explain it away) referred to in condition (2), though this appreciating shouldn't imply that *S* forms the attitude that this data warrants.

Moreover, the notion of unwarrantedness here is to be understood in relation to the degree of conviction of *S*'s average IPC. *S*'s degree of conviction in the proposition that *p* will be unwarranted if it deviates to a noteworthy degree from that which his ICPs would form on the basis of considering the same information that *S* was acquainted with, and deviates in the direction of what *S* wants to be true. Note also that I take it, as I think Mele does also, that we should think of an attitude as being unwarranted relative to, not just the data that *S* possesses, but possibly also to data which *S* didn't possess but which *S* easily could or should have possessed, i.e. data which was "easily available" to her. This is because the biased behaviour referred to in (2) may include selective gathering of further data, and in such cases we should consider the degree of conviction as being unwarranted relative to the data/considerations that she *neglected* to collect/consider, because of this selective evidence search (Davies 2009: 75). Note also that unlike Mele, I am omitting any condition which states that *p* must be false, since as I explained earlier, this is not a necessary condition for self-deception.

Before leaving this section, let me attempt a general explanation of why self-deception involves tension of this sort. Self-deception is a phenomenon of normal, not abnormal psychology. It is, by definition, not a pathological phenomenon, as delusion is. Consequently, it is something that is perpetrated by normal people. It is partly constitutive in turn of the idea of a normal person that they are, in general, intellectually able and rational, and are consequently *not* completely immune to the force of good evidence when they encounter it. For insofar as someone showed himself to be insensitive to reason entirely, to that extent his condition would be considered abnormal and pathological. So it is the fact that self-deception is, by definition, perpetrated by normal people, who are generally sensitive to

the force of good evidence, that explains why it must be characteristic of self-deceivers not to be left entirely unperturbed by the evidence, try as they might to dismiss and ignore it. And it is this which also explains why motivationally biased belief only occurs, as numerous writers have remarked, when the unwelcome evidence falls short of being conclusive, and why it 'evaporates when exposed to the light of overwhelming fact' (Johnson 1997: 118).

Does this mean that it is a conceptual or empirical truth that self-deception involves tension? It is not clear what to say here. On the one hand, we may want to agree with Mele in saying that it is not conceptually *necessary* that it does (1997b: 131). For may it not be possible for self-deceivers on occasion to be entirely successful in explaining away the unwelcome evidence, through their biased behaviour, to their own satisfaction, though this would not make them count as delusional in a pathological sense since they would still be disposed to assent to the unwelcome truth were incontrovertible evidence produced? I can see no reason to deny that this could ever occur, or to claim that if it did occur we should not call them self-deceivers. But on the other hand, tension is associated with self-deception because self-deceivers are generally rational beings who are generally sensitive to the force of good evidence, and that is not a contingent truth about self-deceivers. If conceptual truths are all necessary truths then perhaps it's not a conceptual truth that self-deception involves tension, though this may be too narrow a view of what conceptual truths are.

### *6. Deep Conflict Cases.*

If this is what the tension of self-deception amounts to–feeling harassed by doubts, wavering and uncertainty, and the behaviour associated with such attitudes etc.,–then the above modified deflationism is well suited to account for it. However, traditionalists may accuse us of having mischaracterized the explanandum above. They may insist that the tension of self-

deception is something altogether more, as one might say, 'schizophrenic'. On the understanding they have in mind, this explanation in terms of degrees of conviction might not manage so well. And it does seem that certain philosophers understand this tension in ways that, as I will argue, don't seem compatible with the idea that the self-deceiver really has an unwarranted belief or degree of confidence in the welcome proposition at all.

Some philosophers understand the idea of behavioural tension in a particular way, where they imagine the self-deceiver as indicating with *what she says* that she believes that *p*, perhaps by denying that not-*p* in a defensive, evasive, or flustered fashion, while also indicating with *non-verbal behaviour* that she knows the unwelcome truth, not-*p* (e.g. Audi 1997, Lee 2002: 282, Rorty 1988: 11). The relevant non-verbal behaviour most often mentioned is *avoidance behaviour*, in which the self-deceiver steers clear of things that may remind her of, or insinuate, or put her face-to-face with the truth, behaviour which points strongly to belief in the truth (e.g. Funkhouser 2005, Patten 2003: 241, Pears 1991: 398, Williams 1970/1993: 151). Funkhouser illustrates the structure of a supposedly typical case with the example of a balding man who denies that he's bald and yet ensures that his baldness is kept hidden (using the 'comb-over technique', posing at a certain angle for photographs, refusing to let his wife tussle his hair etc., see 2005: 296) and with another example of a woman called Nichole who avows to her concerned friends and to herself that her husband is not having an affair with a certain other woman in the face of strong evidence to the contrary, but who goes out of her way to avoid places where she would find them together if the reports were true. Funkhouser calls these cases of *deeply conflicted* self-deception (2009). I will adopt this terminology, but without assuming that these are cases of self-deception, by calling them *deep conflict cases*.

These ideas are in stark contrast with the views of others who explicitly deny that self-deception involves such conflict between verbal and non-verbal behaviour. Cosentino,

for instance, says that '[t]he genuine case of self-deception, the one which puzzles us, is *not* where his actions "say" something different than his words but where his actions fit his words, where what is apparent (*p*) to every reasonable person (with the same evidence) seems not to be apparent to the self-deceived person' (Cosentino 1980: 450). Cosentino would rather refer to deep conflict cases as hypocrisy or pretence, and not self-deception. Van Leeuwen also describes what he takes as paradigm cases in which the subject avows that *p* and backs that avowal up with actions including, crucially, taking serious risks on the assumption that *p*, and he rejects these deep conflict cases as ingenuine (Van Leeuwen 2007). Philosophers who reject the idea of this kind of behavioural conflict usually take it to be fundamental that self-deception involves false and unwarranted belief.

Regarding these deep conflict cases, what should we conclude from such behaviour about the person's beliefs? Some deep conflict theorists have thought that this would be evidence that the subject both believes that *p* and believes that not-*p*. Where *p* is the welcome falsehood, they have thought that their verbal assertions that *p* are evidence that they believe that *p*, while the avoidance behaviour would be evidence that they also believe not-*p* (e.g. Rey 1988: 264 & 278). Others have suggested that there may be no determinate answer to the question of what the person believes (e.g. Funkhouser 2009, Hamilton 2000: 25). These responses seem motivated by the assumption that the relevant belief-that-*p*-consistent and belief-that-not-*p*-consistent behaviour are of *equal evidential weight* as indicators of what the subject believes. This is questionable. As Funkhouser notes (2005: 300), we generally take non-verbal behaviour as being privileged over verbal behaviour when it comes to belief attribution, and he takes (in his earlier article, though not his later one) these deep conflict cases as indicative of someone who knows the truth but who does not really believe the contrary falsehood at all, despite the avowal that he does. Funkhouser seems to me to be correct in this, though he wonders why we privilege non-verbal behaviour. The explanation

for this, I believe, can be given in terms of risk-taking, an explanation hinted at by Funkhouser himself (2005: 307), and elaborated more by Gendler.

## *7. Belief and Risk-Taking.*

It is not true that all belief-consistent behaviour is of equal weight as evidence for deciding on what someone believes. What is of paramount importance is how the person acts when he/she understands that there would be something to be gained, or costs to be incurred, if one were to act on the assumption that the belief is true, were it actually not true. That is, the extent to which *S* really believes that *p* can be gauged by observing the risks he/she is willing to take on that assumption. To that extent, not all belief-that-*p*-consistent behaviour is of equal evidential worth, because not all belief-that-*p*-consistent behaviour is associated with equal levels of risk or of possible gain. So for instance, my action of saying that the lake is completely frozen over does not carry the same weight as my action of going out on the lake to skate, other things being equal, when it comes to deciding whether I really believe that, though in the circumstances it may be a perfectly good reason for taking me to believe it. Or my act of saying that the food isn't poisonous doesn't carry the same weight as the act of tasting the food, when deciding whether I believe the food isn't poisonous, other things being equal.

This is relevant to our assessment of the status of the subject's belief in the deep conflict cases purported to be cases of self-deception, and is a point noted by Gendler. Consider the following typical deep conflict case from Gendler (2007: 244-245), where a man who has been diagnosed as terminally ill denies and gives explanations against the diagnosis, suggesting that he believes he is well (*p*). This is belief-that-*p*-consistent behaviour. However, he later comes by the opportunity to take a powerful drug that would

cure that illness, but that would be detrimental to anyone without that illness. He opts to take the drug. This is belief-that-not-*p*-consistent behaviour. Clearly though, both behaviours don't have equal evidential weight. The reason is that his acting as if *p* were true when quizzed by others about his health would not cause him to incur any significant cost or loss if it were really the case that not-*p*. If he were pretending, it would cost nothing to keep up the pretence verbally in these circumstances. But his acting as if *p* were true when he's offered the drug (which would involve not taking it) would cause him to incur a significant loss were it really the case that not-*p*. If the man was pretending that *p* here, maintaining the pretence would be very costly, in that he would miss an important opportunity. This is why the belief-that-not-*p*-consistent behaviour evidentially trumps the belief-that-*p*-consistent behaviour, and we reasonably infer, as Gendler recommends, that he really thought that he was terminally ill after all, and must have been just pretending to others and to himself that things were fine. These points account for why 'actions speak louder than words', that is, why we generally prioritize non-verbal over verbal behaviour when 'inferring' belief, and they allude to classic methods for exposing malingering. Of course, circumstances sometimes do obtain where a lot rides on verbal expressions of belief (think of someone in a game-show with big prizes at stake), but more often risk is associated with non-verbal behaviour.

In the deep conflict cases, as they are typically described, the subject's belief-that-*p*-consistent behaviour would not be at all costly if it were true that not-*p*. It is just verbal behaviour, usually just a matter of saying things in front of people. However, the subjects fail to display belief-that-*p*-consistent behaviour in circumstances where it would be costly to act on that assumption, given that not-*p*. I take it that we would ordinarily consider this good evidence that the person doesn't really believe what they profess to at all. Verbal assertions that *p* do not support a judgment that one believes that *p* where there is an unwillingness on the subject's part to put something on the line on that assumption. Believing that *p* implies a

willingness to, so to speak, 'put one's money where one's mouth is'. But because subjects in these cases are best interpreted as knowing the truth that not-*p*, they could not be accounted for with the modified deflationary theory given above, which demands that they have a degree of confidence in the welcome proposition that *p* which seems incompatible with our saying that they know the contrary unwelcome one. I now wish to argue that with these cases, self-deception has been confused with something else.

## *8. Self-deception and Escapism.*

The idea that self-deceivers may not have a false or unwarranted belief has been around for quite some time. However, it is worth noting that those who maintain it usually do not deny that self-deception may sometimes involve unwarranted, false belief. One of the first philosophers to hold this view may have been M.W. Martin. For Martin, self-deception 'need not involve ignorance and unwarranted belief' (1979: 446) but may instead involve the "intentional evasion of unpleasant topics and truths" (1997: 122), truths which one knows to be true. This 'need not' suggests that Martin thinks self-deception *can* involve ignorance and unwarranted belief. Though Funkhouser initially categorized unwarranted belief cases as 'self-delusion' rather than self-deception, reserving the term 'self-deception' for deep conflict cases (2005), he later renounced this, thinking that his original distinction was *ad hoc* and stipulative. His new view is that deep conflict cases are only *one kind* of self-deception, along with unwarranted belief cases (2009). Gendler takes deep conflict cases to be 'the cleanest and most interesting cases of self-deception' (2007: 233), which again implies that she doesn't deny that the rival false/unwarranted belief cases count as self-deception also. And although some philosophers still deny that false/unwarranted belief cases are self-deception, preferring to call them 'delusion' (see Audi 2007), Mele (2010) has recently done some

'experimental philosophy' or surveys, and has presented evidence that naive subjects are inclined to use 'self-deception' for false/unwarranted belief cases. (Though suggestive, this evidence is not conclusive however; one would want to know much more here, like how they would categorize deep conflict cases, and whether after considering those, they would be inclined to revise their initial decision with the unwarranted belief cases, thinking deep conflict cases as being more deserving of the title "self-deception").

Let's look at these cases as these theorists see them. Such cases as Martin has in mind are ones where people avoid facing up to certain truths or responsibilities. He gives the vivid example of someone who avoids facing up to his own mortality:

> It is a truism that a person can know very well that he will some day die and yet throughout most of his life manage never to honestly confront death. Avoiding such genuine confrontations need not be a matter of keeping oneself ignorant or adhering to false beliefs concerning death and one's responses to it. Primarily it is a matter of pretending there is nothing to be concerned about and withholding one's emotions– refusing to feel rather than refusing to believe. It seems perfectly natural to say of such a person that he is deceiving himself in evading full recognition, and admission to himself, that he will die (1979: 442).

This is a case of the 'intentional evasion of unpleasant topics and truths' (Martin 1997: 122), and it can certainly be done without this causing one to lose one's knowledge that one will die or have a false belief about the matter. It is more a matter of our avoiding reflection on the knowledge that we have. But is this really self-deception?

It seems that now we have on our hands two phenomena that can be distinguished. Roughly, there is 1) avoiding reflecting on and confronting an unpleasant truth that one

knows about, and 2) not believing this unwelcome truth but having an unwarranted level of skepticism towards it due to your desires biasing your evaluation of the issue. Now deep conflict theorists, who regard something like (1) as instantiating self-deception, will if pressed generally not deny the common presumption that (2) counts as self-deception also.

So the question, then, is whether both of these phenomena are in fact self-deception. Is this so? Do (1) and (2) constitute a single psychological kind? It would be *surprising* if this were so, for a number of reasons. First, there is such a world of difference between (1) and (2). The concept of self-deception would be thus ambiguous. For if someone told me, 'Larry, he's deceiving himself about such-and-such', I would not know what to think, and would have to follow up, 'do you mean he doesn't believe that such-and-such is true, or do you mean he knows it but is avoiding it, won't talk about it and admit it, etc.?' Secondly, though we can see similarities between (2) and interpersonal deception, and can hence appreciate why they would both count as species of the one genus (i.e. deception)[64], interpersonal deception and (1) have little or nothing in common, making it a mystery why they would be considered species of the same genus.[65]

Thirdly, and most importantly, these theorists have overlooked the fact that we have another term used for picking out (1): *escapism*. One of the few people in the debate who has shown sensitivity to the distinction between self-deception and escapism, in a paper deserving of more attention than it has received, is John Longeway. Escapism, Longeway suggests, comes in mild and harmless forms, as when we indulge in entertainments which

[64] The similarities are that 1) both cases typically involve the deceived having a false belief, and 2) in both cases, the actions of the deceiver are responsible for the deceived having this false belief, though in self-deception, of course, the deceiver is the same person as the deceived.

[65] Funkhouser anticipates this objection by suggesting that since conceiving of self-deception on the model of the interpersonal case leads to well-known problems, we shouldn't feel that there needs to be a close similarity between self-deception and the interpersonal case (see 2005: 304 & 299). But surely there must be *some* similarity, *some* shared features between the two cases to sustain the point that they are both species of deception, and deep conflict cases simply don't appear to have the requisite similarity.

temporarily draw us away from our everyday troubles. But sometimes, he remarks, 'we speak of a more serious escapism, in which we avoid thinking about what we know to be so, not in the course of recreation or to keep unpleasant thought out of mind as long as they are not necessary, but as a defence against reality itself' (1990: 1). Such escapism 'attempts to keep beliefs one does not like out of consciousness...and should they enter consciousness, to distract one from them or put them out of mind' (1990: 2). These activities, he claims, exercise us quite often, though they don't attract the term 'escapism' unless we habitually try to avoid the reality when we should or need to consider or face up to it. Examples of escapist techniques mentioned by Longeway include distracting oneself with irrelevant concerns to force the belief out of consciousness, denying verbally or pretending to oneself or others not to hold it, avoiding situations which would remind one of the matter, and restricting one's company to those who will not remind one of it (1990: 1), all of which are activities deep conflict theorists have associated with self-deception. Things used for escapist purposes can be anything from one's own imagination, to alcohol and drugs. Cases of escapism are ones where we say that someone 'puts to the back of his mind' what he 'really knows deep down'. Such people may verbally deny and counter the truth, but this behaviour may be explained as, not an expression of belief, but as being part of their attempt to ignore the reality, and divert attention/conversation to something else. Deep conflict cases therefore resemble paradigm cases of escapism, and thus Martin, Gendler, and Funkhouser may be open to the charge of failing to respect the distinction between an escapist and a self-deceiver.

If (1) and (2) are indeed distinct phenomena, why have they been often conflated in the literature? There are a number of possible explanations for this. Longeway says that many of the techniques of escapism may result in self-deception too, that is, they may result in the elimination or prevention of a warranted belief rather than in avoiding awareness of it, and

that is why the two phenomena, self-deception and escapism, can easily be confused with one another.

Furthermore, because escapism is related to self-deception in a number of ways (for instance, both occur when unwelcome truths are pressing for recognition) they may often be found, in real life cases, as bound up with each other. The alcoholic, for instance, who might use alcohol as a means of escape from an unpleasant reality, might have also deceived himself into believing that he can indefinitely avoid the unpleasant reality, or that his present behaviour is sustainable, or that his lifestyle is really not so bad, or that he's a victim, or that he doesn't have it in him to change. It may not be easy to separate or distinguish the elements of escapism and self-deception in such a case. But the fact that both phenomena might often be found together doesn't undermine their conceptual distinctness.

It may, in addition, be pointed out that part of what motivates the thought that these deep conflict cases are self-deception is the idea that self-deception must be closely modeled on interpersonal deception. For typically in interpersonal deception, the deceiver does know the truth, so one might expect that the self-deceiver must know the truth also, just as the subject's apparently do in these deep conflict cases. As has been pointed out before, however, this way of reasoning is fallacious.

Again, this all presupposes that the best way to interpret the subject's behaviour in these deep conflict cases is as indicating that they really know the truth. From the descriptions given of such cases in the literature, this presumption seems justified, but whether this is the best explanation of any case ultimately depends on its details and how it is described. On other descriptions, the behaviour may not weigh so heavily towards that conclusion, indicating some degree of confidence in the welcome proposition (evidenced by a willingness to take some risks on the assumption that *p*, for instance), and so the case may be amenable to treatment with the above modified deflationary account if this confidence goes

beyond what's warranted. The important thing is just to bear in mind what distinguishes escapism from self-deception, which is that the former involves knowing/believing ('deep down') an unwelcome truth but avoiding facing up to it, while the latter would require that the subject is skeptical towards this proposition to some unwarranted degree.

In conclusion, if the claim that self-deception involves tension is to be understood as meaning that the self-deceiver is usually afflicted with doubts about what he wants to be true and exhibits behaviour which displays something short of wholehearted commitment to the welcome assumption, then this idea can easily be assimilated into a deflationary account which frames things in terms of unwarranted levels of confidence rather than unwarranted belief in the welcome proposition. If, however, the idea is to be understood as meaning that the self-deceiver displays behaviour which is more characteristic of one who really knows the unwelcome truth, then we are under no obligation to accommodate such a thing in our theory, since it is the mark of an evader and an escapist, and not a self-deceiver.

# Bibliography

(2006) *American Heritage Dictionary*, 4th Edition, Houghton Mifflin Harcourt.

American Psychiatric Association, (2000) *Diagnostic and Statistical Manual of Mental Disorders*, 4th Edition, text revision, American Psychiatric Association.

Anderson, M.C. & Green, C. (2001) 'Suppressing Unwanted Memories by Executive Control', *Nature*, 410(6860), 366-369.

Anscombe, G.E.M. (1966) *Intention*, Oxford: Basil Blackwell.

Audi, R. (2007) 'Belief, Intention, and Reasons for Action', in M. Timmons, J. Greco & A. Mele (eds.), *Rationality and the Good*, New York: Oxford University Press, pp.248-262.

– (1997) 'Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele', *Behavioral and Brain Sciences*, 20, 104.

– (1989) 'Self-deception and Practical Reasoning', *Canadian Journal of Philosophy*, 19, 247-266.

– (1982) 'Self-Deception, Action, and Will', *Erkenntnis*, 18, 133-158.

– (1976) 'Epistemic Disavowals and Self-Deception', *The Personalist*, 57, 378-385.

Austin, J.L. (1966) 'Three Ways of Spilling Ink', *Philosophical Review*, 75, 427-440.

Bach, K. (2009) 'Self-Deception', in B.P. McLaughlin, Ansgar Beckermann, & Sven Walter (eds.) *The Oxford Handbook of Philosophy of Mind*, Oxford: Clarendon Press, pp.781-796.

– (1997) 'Thinking and Believing in Self-Deception', *Behavioral and Brain Sciences*, 20, 105.

– (1981) 'An Analysis of Self-Deception' in *Philosophy and Phenomenological Research*, 41, 351-370.

Balcetis, E. (2008) 'Where the Motivation Resides and Self-Deception Hides: How Motivated Cognition Accomplishes Self-Deception', *Social and Personality Psychology Compass* 2, 361-381.

Barnes, A. (1997) *Seeing Through Self-Deception*, Cambridge: Cambridge University Press.

Baumeister, R.F. (1993) 'Lying to Yourself: The Enigma of Self-Deception', in M. Lewis & C. Saarni (eds.), *Lying and deception in everyday life*, New York: Guilford Press, pp.166-183.

Bennett, M.R. & Hacker, P.M.S. (2003) *Philosophical Foundations of Neuroscience*, Oxford: Blackwell.

Bermúdez, J.L. (2000) 'Self-Deception, Intentions, and Contradictory Beliefs', *Analysis*, 60, 309-319.

– (1997) 'Defending Intentionalist Accounts of Self-Deception', *Behavioral and Brain Sciences*, 20, 107-108.

Bird, A. (1994) 'Rationality and the Structure of Self-Deception', in *European Review of Philosophy, vol.1*. Stanford: CSLI Publications, pp.19-38.

Bok, S. (1980) 'The Self Deceived', *Social Science Information*, 19, 923-935.

Borge, S. (2003) 'The Myth of Self-Deception', *Southern Journal of Philosophy*, 41, 1-28.

Boyd, I.D, (2006, revised 2010) 'Self-Deception', *Stanford Encyclopedia of Philosophy*.

Bratman, M. (1984) 'Two Faces of Intention', *Philosophical Review*, 93, 375-405.

Bulevich, J.B. Roediger H.L. Balota, D.A. & Butler A.C. (2006) 'Failures to Find Suppression of Episodic Memories in the Think/No-Think Paradigm', *Memory and Cognition*, 34, 1569-1577.

Campbell, J. (2001) 'Rationality, Meaning, and the Analysis of Delusion', *Philosophy, Psychology, Psychiatry*, 8, 89-100.

Champlin, T.S. (1988) *Reflexive Paradoxes*, London; New York: Routledge.

– (1977) 'Self-Deception: A Reflexive Dilemma', *Philosophy*, 52, 281-299.

Chisholm, R.M & Feehan, T.D. (1977) 'The Intent to Deceive', *Journal of Philosophy*, 74, 143-159.

Cosentino, D.A. (1980) 'Self-Deception Without Paradox', *Philosophy Research Archives*, 6, 444-465.

da Costa, N.C.A. & French, S. (1990) 'Belief, Contradiction and the Logic of Self-Deception', *American Philosophical Quarterly*, 27, 179-197.

Davidson, D. (2004/1997) 'Who is Fooled?', in *Problems of Rationality*, Oxford: Clarendon Press, pp.213-230.

– (2004/1986) 'Deception and Division', in *Problems of Rationality*, Oxford: Clarendon Press, pp.199-212.

– (2004/1985) 'Incoherence and Irrationality', in *Problems of Rationality*, Oxford: Clarendon Press, pp.189-197.

– (2004/1982) 'Paradoxes of Irrationality', in *Problems of Rationality*, Oxford: Clarendon Press, pp.169-187.

– (2001/1978) ‘Intending’, in *Essays on Actions and Events*, Oxford: Oxford University Press, pp.82-102.

– (1999) ‘Reply to Jon Elster’, in L.E. Hahn (ed.), *The Philosophy of Donald Davidson*, Chicago, Illinois: Open Court, pp.443-445.

Davies, M. (2009) ‘Delusion and Motivationally Biased Belief: Self-Deception in the Two-Factor Framework’, in T. Bayne & J. Fernandez (eds.), *Delusions and Self-Deception: Affective Influences on Belief Formation*, Hove: Psychology Press, pp.71-86.

Demos, R. (1960) ‘Lying to Oneself’, *Journal of Philosophy* 57: 588-595.

Ditto, P.H. Munro, G.D. Apanovitch, A.M. Scepansky, J.A. & Lockhart, L.K. (2003) ‘Spontaneous Skepticism: The Interplay of Motivation and Expectation in Responses to Favourable and Unfavourable Medical Diagnoses’, *Personality and Social Psychology Bulletin*, 29, 1120-1132.

Ditto, P.H. Scepansky, J.A. Munro, G.D. Apanovitch, A.M. & Lockhart, L. (1998) ‘Motivated Sensitivity to Preference-Inconsistent Information’, *Journal of Personality and Social Psychology*, 75, 53-69.

Ditto P.H. & Lopez D.F. (1992) ‘Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions’, *Journal of Personality and Social Psychology*, 63, 568-584.

Donnellan, K.S. (1963) 'Knowing What I am Doing', *Journal of Philosophy*, 60, 401-409.

Doucet, M. (forthcoming) 'Can We Be Self-Deceived about What We Believe? Self-Knowledge, Self-Deception, and Rational Agency', *European Journal of Philosophy*.

Dunning, D. (1999) 'A Newer Look: Motivated Social Cognition and the Schematic Representation of Social Concepts', *Psychological Inquiry*, 10, 1-11.

Elster, J. (1999) 'Davidson on Weakness of Will and Self-Deception' in L.E. Hahn (ed.), *The Philosophy of Donald Davidson*, Chicago and La Salle: Open Court, pp.425-442.

Eriksson, L. & Hájek, A. (2007) 'What are Degrees of Belief?', *Studia Logica*, 86, 183-213.

Fingarette, H. (1998) 'Self-Deception Needs No Explaining', *Philosophical Quarterly*, 48, 289-301.

Foley, R. (2009) 'Beliefs, Degrees of Belief, and the Lockean Thesis', in F. Huber & C. Schmidt-Petri (eds.) *Degrees of Belief*, Dordrecht; London: Springer, pp.37-47.

– (1986), 'Is it Possible to have Contradictory Beliefs?', *Midwest Studies in Philosophy*, 10, 327-355.

Frey, D. & Stahlberg, D. (1986) 'Selection of Information after Receiving More of Less Reliable Self-Threatening Information', *Personality and Social Psychology Bulletin*, 12, 434-441.

Funkhouser, E. (2009) 'Self-Deception and the Limits of Folk Psychology', *Social Theory and Practice*, 35, 1-13.

– (2005) 'Do The Self-Deceived Get What They Want?' *Pacific Philosophical Quarterly*, 86: 295-312.

Gardiner, P. (1969-70) 'Error, Faith and Self-Deception', *Proceedings of the Aristotelian Society*, 70, 221-243.

Gardner, S. (1993) *Irrationality and the Philosophy of Psychoanalysis*, Cambridge: Cambridge University Press.

Gendler. T.S. (2007) 'Self-Deception as Pretense', *Philosophical Perspectives*, 21: 231-258.

Geraerts, E. Merckelbach, H. Jelicic, M. & Smeets, E. (2006) 'Long-Term Consequences of Suppression of Intrusive Anxious Thoughts and Repressive Coping', *Behaviour Research and Therapy*, 44, 1451-1460.

Gorr, M. & Horgan T. (1982) 'Intentional and Unintentional Actions', *Philosophical Studies*, 41 251-262.

Gozzano, S. (1999) 'Davidson on Rationality and Irrationality', in M. De Caro (ed.) *Interpretations and Causes: New Perspectives on Donald Davidson's Philosophy*, Dordrecht; Boston; London: Kluwer Academic Publishers, pp.137-149.

Graham, G. (1986) 'Russell's Deceptive Desires', *The Philosophical Quarterly*, 36, 223-229.

Gur, R. & Sackeim, H. (1979) 'Self-Deception: A Concept in Search of a Phenomenon', *Journal of Personality and Social Psychology*, 37, 147-169

Gustafson, D. (1975) 'The Range of Intentions', *Inquiry*, 18, 83-95.

Haight, M.R. (1985) 'Tales from a Black Box', in M.W. Martin (ed.), *Self-Deception and Self-Understanding*, Lawrence, Kansas: University Press of Kansas, pp.244-260.

Hales, S.D. (1994) 'Self-Deception and Belief Attribution', *Synthese*, 101, 273-289.

Hamilton, A. (2000) 'The Authority of Avowals and the Concept of Belief', *European Journal of Philosophy*, 8, 20-39.

Hamlyn, D.W. (1971) 'Self-Deception', *Proceedings of the Aristotelian Society* (Supplementary Volume 35), 45-60.

Hampshire, S. (1970/1959) *Thought and Action*, London: Chatto and Windus.

Hellman, N. (1983) 'Bach on Self-Deception', *Philosophy and Phenomenological Research*, 44, 113-120.

Holton, R. (2001) 'What is the Role of the Self in Self-Deception?', *Proceedings of the Aristotelian Society*, 101, 53-69.

Holton, B. & Pyszczynski, T., (1989) 'Biased Information Search in the Interpersonal Domain', *Personality and Social Psychology Bulletin*, 15, 42-51.

Hunter, D. (1996) 'On the Relation between Categorical and Probabilistic Belief', *Noûs*, 30, 75-98.

Johnson, E.A. (1997) 'Real ascriptions of self-deception are fallible moral judgments', *Behavioral and Brain Sciences*, 20, 117-118.

Johnston, M. (1988) 'Self-Deception and the Nature of Mind', in B.P. McLaughlin & A.O. Rorty (eds.) *Perspectives on Self-Deception*, Berkeley etc.: University of California Press, pp.63-91.

Kipp, D. (1980) 'On Self-Deception', *The Philosophical Quarterly*, 30, 305-317.

Kunda, Z. (1999) *Social Cognition: Making Sense of People*, Cambridge, Mass.: MIT Press.

– (1990) 'The Case for Motivated Reasoning', *Psychological Bulletin*, 108, 480-498.

– (1987) 'Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories', *Journal of Personality and Social Psychology*, 53, 636-647.

Kunda, Z. Fong, G.T. Sanitioso, R. & Reber, E. (1993) 'Directional Questions Direct Self-Conceptions', *Journal of Experimental Social Psychology*, 29, 63-86.

Lazar, A. (1999) 'Deceiving Oneself of Self-Deceived? On the Formation of Beliefs "Under the Influence"', *Mind*, 108, 265-290.

Lee, B.D. 'Shoemaker on Second-Order Belief and Self-Deception', *Dialogue*, 41, 279-289.

Liberman, A. & Chaiken, S. (1992) 'Defensive Processing of Personally Relevant Health Messages', *Personality and Social Psychology Bulletin*, 18, 669-679.

Lockie, R. (2003) 'Depth-Psychology and Self-Deception', *Philosophical Psychology*, 16, 127-148.

Longeway, J.L. (1990) 'The Rationality of Escapism and Self-Deception', *Behavior and Philosophy*, 18, 1-19.

Losonsky, M. (1997) 'Self-Deceivers' Intentions and Possessions', *Behavioral and Brain Sciences* 20, 121-122.

Lundgren, S.R. & Prislin, R. (1998) 'Motivated Cognitive Processing and Attitude Change', *Personality and Social Psychology Bulletin*, 24, 715-726.

Malle B.F. & Knobe, J. (1997) 'The Folk Concept of Intentionality', *Journal of Experimental Social Psychology*, 33, 101-121.

Martin, M.W. (1997) 'Self-Deceiving Intentions', *Behavioral and Brain Sciences*, 20, 122-123.

– (1986) *Self-Deception and Morality*, Kansas: University of Kansas Press.

– (1979) 'Self-Pretence, and Emotional Detachment', *Mind*, 88, 441-446.

Martin, T. (1998) 'Self-Deception and Intentional Forgetting: A Reply to Whisner', *Philosophia* 26, 181-194.

McLaughlin, B.P. (1996) 'On The Very Possibility Of Self-Deception' in R.T. Ames & W. Dissanayake (eds.), *Self and Deception: A Cross-Cultural Philosophical Enquiry*, Albany: State University of New York, pp.31-51.

– (1988) 'Exploring the Possibility of Self-Deception in Belief' in B.P. McLaughlin & A.O. Rorty (eds.) *Perspectives on Self-Deception*, Berkeley etc.: University of California Press, pp.29-62.

Mele, A.R. (2010) 'Approaching Self-Deception: How Robert Audi and I part company', *Consciousness and Cognition*, 19, 745-750.

– (2007) 'Self-deception and Three Psychiatric Delusions', in M. Timmons, J. Greco & A. Mele (eds.), *Rationality and the Good*, New York: Oxford University Press, pp.163-175.

– (2004) ‘Motivated Irrationality’, in A.F. Mele & P. Rawling (Eds.) *The Oxford Handbook of Rationality*, Oxford; New York: Oxford University Press, pp.240-256.

– (2003) *Motivation and Agency*, New York; Oxford University Press.

– (2001) *Self-Deception Unmasked*, Princeton; Oxford: Princeton University Press.

– (1998) ‘Motivated Belief and Agency’, *Philosophical Psychology*, 11, 353-369.

– (1997a) ‘Real Self-Deception’, *Behavioral and Brain Sciences*, 20, 91-102.

– (1997b) ‘Understanding and Explaining Real Self-Deception’, *Behavioral and Brain Sciences*, 20, 127-134.

– (1987) ‘Recent Work on Self-Deception’, *American Philosophical Quarterly*, 24, 1-17.

– (1983) ‘Self-Deception’, *The Philosophical Quarterly*, 33, 365-377.

– (1982) ‘“Self-Deception, Action, and Will”: Comments’, *Erkenntnis*, 18, 159-164.

Mele, A.R. & Moser, P.K. (1994) ‘Intentional action’, *Noûs*, 28, 39-68.

Mellor, D.H. (1977-78) ‘Conscious Belief’, *Proceedings of the Aristotelian Society*, vol.78, 87-101.

*Memento* (2000), directed by Christopher Nolan.

Michel, C. & Newen, A. (2010) 'Self-Deception as Pseudo-Rational Regulation of Belief', *Consciousness and Cognition*, 19, 731-744.

Miller, A.R. (1980) 'Wanting, Intending, and Knowing What One is Doing', *Philosophy and Phenomenological Research*, 40, 334-343.

Moran, R. (2001) *Authority and Estrangement*, Princeton; Oxford: Princeton University Press.

Nelkin, D. (2002) 'Self-Deception, Motivation, and the Desire to Believe', *Pacific Philosophical Quarterly* 83: 384-406.

Newman, L.S. (1999) 'Motivated Cognition and Self-Deception', *Psychological Inquiry*, vol.10, 59-63.

Noordhof, P. (2009) 'The Essential Instability of Self-Deception', *Social Theory and Practice*, 35, 45-71.

Pataki, T. (1997) 'Self-Deception and Wish-Fulfilment', *Philosophia*, 25, 297-322.

Patten, D. (2003) 'How do we Deceive Ourselves?', *Philosophical Psychology*, 16, 229-246.

Pears, D. (1991) 'Self-Deceptive Belief-Formation', *Synthese*, 89, 393-405.

– (1984) *Motivated Irrationality*, Oxford: Clarendon Press.

Perring, C. (1997) 'Direct, Fully Intentional Self-Deception is Also Real', *Behavioral and Brain Sciences*, 20, 123-124.

Poellner, P. (2004) 'Self-Deception, Consciousness and Value: The Nietzschean Contribution', *Journal of Consciousness Studies*, 11, 44-65.

Pugmire, D.R. (1969) '"Strong" Self-Deception', *Inquiry*, 17, 339-361.

Pyszczynski, T. Greenberg, J. & Holt, K. (1985) 'Maintaining Consistency between Self-Serving Beliefs and Available Data: A Bias in Information Evaluation', *Personality and Social Psychology Bulletin*, 11, 179-190.

*Raging Bull* (1980), directed by Martin Scorsese.

Rassin, E. (2005) *Thought Suppression*, Amsterdam; London: Elsevier.

Rassin, E. Merckelbach, H. & Muris, P. (1997) 'Effects of Thought Suppression on Episodic Memory', *Behaviour Research and Therapy*, 35, 1035-1038.

Reilly, R. (1976) 'Self-Deception: Resolving the Epistemic Paradox', *The Personalist*, 57, 391-394.

Rey, G. (1988) 'Towards a Computational Account of Akrasia and Self-Deception', in B.P. McLaughlin & A.O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley etc.: University of California Press, pp. 264-296.

Rorty, A.O. (1998) 'The Deceptive Self: Liars, Layers, and Lairs', in B.P. McLaughlin & A.O. Rorty (eds.), *Perspectives on Self-Deception*, Berkeley etc.: University of California Press, pp.11-28.

Ross, G. (1982) 'Knowledge and Intentional Action', *Philosophical Studies*, 41, 263-266.

Sanitioso, R. Fong, G.T. & Kunda, Z. (1990) 'Motivated Recruitment of Autobiographical Memories', *Journal of Personality and Social Psychology*, 59, 229-241.

Sarbin, T.R. (1981) 'On Self-Deception', *Annals of the New York Academy of Sciences*, 364, 220-235.

Sartre, J.P. (1956) *Being and Nothingness*, trans. Hazel Barnes, New York: Philosophical Library.

Searle, J.R. (1995) *The Rediscovery of the Mind,* Cambridge, Mass.; London: MIT Press.

Scott-Kakures, D. (2002) 'At "Permanent Risk": Reasoning and Self-Knowledge in Self-Deception' in *Philosophy and Phenomenological Research*, 65, 577-603.

Scott-Kakures, D. (1996) 'Self-Deception and Internal Irrationality', *Philosophy and Phenomenological Research*, 56: 31-56.

Siegler, F.A. (1963) 'Self-Deception', *Australian Journal of Philosophy*, 41, 29-43.

Soteriou, M. (2009) 'Introduction', in M. Soteriou & L. O'Brien (Eds.) Mental Actions, Oxford: Oxford University Press, pp. 1-16.

Steffen, L.H. (1986) *Self-Deception and the Common Life*, New York etc.: Peter Lang.

Svenson, O. (1981) 'Are We All Less Risky And More Skilful Than Our Fellow Drivers?', *Acta Psychologica*, 47, 143-148.

Szabados, B. (1974) 'Self-Deception', *Canadian Journal of Philosophy*, 4, 51-68.

– (1973) 'Wishful Thinking and Self-Deception', *Analysis*, 33, 201-205.

Talbott, W.J. (1995) 'Intentional Self-Deception in a Single Coherent Self', *Philosophy and Phenomenological Research*, 55, 27-74.

Van Leeuwen, D.S.N. (2008) 'Finite Rational Self-Deceivers', *Philosophical Studies*, 139, 191-208.

– (2007), 'The Product of Self-Deception', *Erkenntnis*, 67, 419-437.

Walker, W.R. Vogl, R.J. & Thompson, C.P. (1997) 'Autobiographical Memory: Unpleasantness Fades Faster Than Pleasantness Over Time', *Applied Cognitive Psychology*, 11, 399-413.

Wason, PC. (1960) 'On the Failure to Eliminate Hypotheses in a Conceptual Task', *The Quarterly Journal of Experimental Psycholog*y, 12, 129-140.

Wegner, D.M. (1994/1989) *White Bears and Other Unwanted Thoughts*, New York; London: The Guilford Press.

Wegner, D.M. Quillian, F. & Houston, C. (1996) 'Memories Out of Order: Thought Suppression and the Disturbance of Sequence Memory', *Journal of Personality and Social Psychology*, 71, 680-691.

Wegner, D.M. & Erber, R. (1992) 'The Hyperaccessibility of Suppressed Thoughts', *Journal of Personality and Social Psychology*, 63, 903-912.

Weinstein, N.D. (1982), 'Unrealistic Optimism about Susceptibility to Health Problems', *Journal of Behavioral Medicine*, 5(4), 441-460.

Weinstein, N.D. (1980) 'Unrealistic Optimism about Future Life Events', *Journal of Personality of Social Psychology*, 39, 806-820.

Whisner, W. (1998) 'A Further Explanation and Defense of the New Model of Self-Deception: A Reply to Martin', *Philosophia*, 26, 195-206.

Williams, B. (1970/1993) 'Deciding to Believe', in *Problems of the Self: Philosophical Papers, 1956-1972*, London: Cambridge University Press, pp.136-151.

Wyer, R.S. & Frey, D. (1983), 'The Effects of Feedback about Self and Others on the Recall and Judgments of Feedback-Relevant Information', *Journal of Experimental Social Psychology*, 19, 540-559.

Yanal, R.J. (2007) 'Self-Deception and the Experience of Fiction', *Ratio*, 20, 108-121.