AN ANALYSIS OF THE POSSIBILITY OF DELIBERATE SELF-DECEPTION

by
Artis Svece

A thesis submitted to the
School of Graduate Studies
in partial fulfilment of the
requirements for the degree of
Master of Arts

Department of Philosophy
Sir Wilfred Grenfell College
Memorial University of Newfoundland

July 1996

Corner Brook Newfoundland

ABSTRACT

The purpose of my thesis is to analyse the possibility of deliberate self-deception. The analysis is developed in two stages. First, I provide an analysis of the term 'self-deception' and the problems that this term presents for philosophers. Second, I analyse the possibility of deliberately making oneself believe what one knows is false.

Even a superficial glance over the literature on self-deception reveals the variety of interpretations the term 'self-deception' has. The differences between philosophers' interpretations of the term makes it hard to understand what can and what cannot be called 'self-deception.' In order to analyse the possibility of deliberate self-deception, I must know what self-deception is. The analysis of the term 'self-deception' turns out to be a rather painstaking enterprise, and I have to separate out the several meanings the word has. In the end of the analysis, I present two meanings of 'self-deception' that the term has in ordinary language, as well as explain the diversity of interpretations the concept has in philosophical discourse.

When I have reached understanding of what the term 'self-deception' means in ordinary language and how it is used by different philosophers, I proceed to the analysis of the possibility of deliberate self-deception. I use the notion of deliberateness in order to distinguish between previously intended and intentional actions which may not be intended beforehand. Deliberate self-deception is an intended action of making oneself believe what one knows is false. The analysis of the possibility of deliberate self-deception is meant to demonstrate the extent to which one can control what one believes. The possibility of such control provides a basis for claiming that the self-deceivers make themselves believe what they know is false. The conclusion of my analysis is that deliberate self-deception is possible, but it is possible only in certain circumstances without which any attempt to deceive oneself fails. The basic methods for such deception are forgetting what one knows and reinterpreting evidence for one's beliefs.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Professors Gunars Tomsons and Sandra Tomsons for their help in preparing this thesis as well as in everything that concerned my life here in Canada. I am grateful for their invaluable advise in every aspect of the philosophical tradition that I had to learn from scratch. I would like to thank them also for their constructive criticism of the ideas I tried to present in my thesis and their patience in correcting my English.

I want to thank also Memorial University of Newfoundland for financial support in the form of Graduate Fellowship without which I would not have been able to study here, and I would like to express my gratitude to Katy Bindon, the Principal of Sir Wilfred Grenfell College, for making my studies at the College pleasant and comfortable.

TABLE OF CONTENTS

1.0 INTRODUCTION

Questions like 'What is self-deception?' and 'How is self-deception possible?' form the core of what could be called the problem of self-deception. Philosophers have discovered that self-deception presents a challenge for many beliefs about the nature of mind, the self, and rationality. Naturally, philosophers have adopted different stances towards self-deception: some are claiming that self-deception does not present any challenge at all, others -- that self-deception provides an important insight into the nature of mind. Because of the lively discussion that the problem of self-deception has generated, self-deception forms a distinct area of interest for the philosophy of mind and the philosophy of psychology.

When I chose to write my thesis on self-deception, the question which I intended to answer was whether it is possible to make myself believe something that I am disposed not to believe. The formulation of the question was also the formulation of my understanding of what self-deception is. The question about the possibility to control, or manipulate, one's beliefs may seem strange, but in fact there is a certain philosophical tradition behind it. The requirement of controlling one's mind has been known since the times of Stoics and Buddhists, and the ethics of self-control has had its adherents ever since. A peculiar version of this ethical tradition is depicted in *Either/Or* by Soren Kierkegaard. In one of the chapters of the book, "Rotation of Crops," Kierkegaard presents methods of how to make one's life interesting even under the most boring conditions. One of the methods is the art of forgetting and remembering (293). If one knows how to remember something in a way one wants and to forget everything one wants, one is "able to play shuttlecock with all existence" (294). While this art concerns only forgetting and remembering, it seems to imply that one can believe whatever one wants. And I wanted to know to what extent, if any, one could control one's beliefs.

As one can see, the context of the problem is very different from what philosophers usually do in the philosophy of mind. Nevertheless, there is a definite connection between the possibility of controlling one's beliefs and the problem of self-deception. By answering the question of whether it is possible to deceive oneself intentionally and be aware of one's intention, I would also answer the question about the possibility controlling one's beliefs.

Since there is extensive literature on self-deception in English, I wanted to know what possibilities for controlling one's beliefs are presented by those who analyse self-deception.

The first discovery I made was the fact that the notion of self-deception is very ambiguous and constituted a problem by itself. So my initial interest in how to make myself believe what I am disposed not to believe might or might not be identified by other philosophers as a characteristic of self-deception. The number of different definitions of self-deception is surprisingly large. Some philosophers consider the meaning of the concept a settled matter and do not formulate their own understanding of self-deception; but if one looks at the definitions that are provided, very rarely one finds two philosophers that have identical interpretations of the concept. Hence, in order to understand what is meant by the term 'self-deception' and whether this meaning is compatible with my interest in the possibility of controlling one's beliefs, I had to answer the question 'What is self-deception?'

I try to answer the question of what self-deception is in Chapter 2 of my thesis. In order to have some criteria for a comparison of the different definitions, I want to concentrate my attention on the meaning of the term 'self-deception' in ordinary language. Analysing different interpretations of the concept, I am first of all answer the question whether these interpretations could represent the meaning the term 'self-deception' has in ordinary language. When the answer is 'no,' I explain why philosophers still use the ordinary language term 'self-deception' for their purposes and what is the connection between their understanding of 'self-deception' and the meaning of 'self-deception' as it is used in ordinary language.

First, I am concentrating on the method of defining self-deception suggested by Raphael Demos. Demos interprets self-deception as being similar to interpersonal deception, the only difference being that the former involves one instead of two persons. The result is a paradoxical formulation of self-deception, namely, making oneself believe what one knows is false. Demos' approach seems very natural, because the word

'deception' does appear in the term 'self-deception' and 'deception' usually refers to a situation where one person deceives another. My analysis of Demos' method of defining 'self-deception' as an interpersonal deception that occurs in one person shows that this method cannot reveal the meaning the concept has in ordinary language. This conclusion applies not only to Demos but also to all those who analyse the meaning of the word 'deception' in order to derive from it the meaning of the term 'self-deception.'

At the same time, I have to admit that the term 'self-deception' has a paradoxical meaning in ordinary language and Demos' interpretation has presented this meaning quite well. I define the first meaning of the ordinary language term 'self-deception' as making oneself believe what one knows is false. I also have to admit that the meaning of 'self-deception' is not always paradoxical. Since I objected to Demos' method of defining the term 'self-deception,' the only way to discern other meanings of 'self-deception' is to describe how the word is used in ordinary language. The second part of Chapter 2 is meant to describe the non-paradoxical meanings the term 'self-deception' has in ordinary language.

One of the occasions when philosophers analyse the ordinary language meaning of 'self-deception' is the discussion of Canfield and Gustavson's and Siegler's interpretations of 'self-deception.' Slightly simplifying their interpretation, I can say that they claim that 'self-deception' means nothing more than an unwarranted belief. In a way I defend this position, but only to claim that this definition depicts the usage of the expression 'to deceive oneself' and not the usage of 'self-deception.' At the same time, I try to show that there is a meaning of 'self-deception' that resembles the one of 'deceiving oneself.' The second meaning that the term 'self-deception' has in ordinary language can be defined as certain state of mind where the false belief is caused by a bias of evidence. Still, the various uses of 'self-deception' I have described so far are just some of the interpretations philosophers have provided. Moreover, most of these interpretations do not present the meaning that the term has in ordinary language. So I continue with my explanation of why so many interpreta-

tions do not fit any of the meanings that 'self-deception' has in ordinary language. I explore the idea that philosophers are trying to explain certain behaviour that is usually associated with the term 'self-deception.'

In Chapter 3 I return to my initial question about the possibility of controlling one's own beliefs. After I have analysed the meaning of the term 'self-deception,' I can indicate how my problem of self-control fits into the problem of self-deception. Some interpretations of self-deception imply that the self-deceiver intentionally brings about his or her beliefs. The suggestion that one controls one's beliefs in self-deception represents the most extreme version of such intentional formation of beliefs. I call this extreme version 'deliberate self-deception,' intending to refer to self-deception that is brought about by a conscious intention to make oneself believe what one knows is false or an intention to make oneself believe what one wants to believe.

I examine three possible forms of deliberate self-deception. All of them are discussed by other philosophers. First of all, I concentrate on the possibility of making oneself 'just like that' believe what one knows is false, namely, the possibility of self-deception as basic action, realization of which does not require any additional intentional actions. Though usually one would deny that such an act is possible, it seems that sometimes the possibility of such a basic action is suggested by people. I will try to present some reasons why such actions are impossible and why one cannot make oneself believe what one knows is false.

In the remaining part of my thesis I address the question of the possibility of deliberate self-deception that is not basic action, that is to say, deliberate self-deception that is realized by means of some additional actions. I am concentrating on two types of actions. The first type could be characterized as making oneself forget what one knows. The second type -- as reinterpreting of the evidence one knows. My conclusion is that it is possible to make oneself believe what one knows is false. For such self-deception to be possible one first of all has to deliberately undermine one's knowledge of the falsity of the

belief, because one cannot consciously hold both the belief that *p* and knowledge that not-*p*. The success of deliberate self-deception never depends completely on the intention of the self-deceiver, and deliberate self-deception is possible only in particular circumstances.

2.0 WHAT IS SELF-DECEPTION?

I dare to say that anybody who decides to enlighten oneself on the subject of self-deception and wants to do this by reading philosophers will quickly loose any comprehension of what self-deception is. Philosophers quite often undermine our everyday understanding of concepts and phenomena, but in the case of self-deception the feeling of confusion is caused by the great variety of interpretations. For example, Stanley Paluch says that a person *X* is self-deceived when:

> (1) *X* believes *p* and *p* is false. (2) *X* knows the evidence which counts against the truth of *p*. (3) *X* has some motive for discounting the evidence. (4) If the motive were lacking *X* would see that *p* is false and its negation true. (5) If the motive were made clear to *X* he would see that it provided no legitimate grounds for his belief. (6) *X* is free to discern the character of his motive (276).

According to Frederick Siegler, if White says to Brown that Brown is deceiving himself, "White is telling Brown that he has an erroneous belief, and he is implying that it is unreasonable for Brown to have such a belief" (473). Herbert Fingarette thinks that "the self-deceiver is one who is in some way engaged in the world but who disavows the engagement, who will not acknowledge it even to himself as his" ("Self-Deception" 81). Robert Audi claims:

> A person, *S*, is in a state of self-deception with respect to a proposition, *p*, if and only if: (1) *S* unconsciously knows that not-*p* (or has reason to believe, and unconsciously and truly believes, that not-*p*); (2) *S* sincerely avows, or is disposed to avow sincerely, that *p*; and (3) *S* has at least one want that explains, in part, both why *S*'s belief that not-*p* is unconscious and why *S* is disposed to avow that *p*, even when presented with what he sees is evidence against *p* (94).

John V. Canfield and Don F. Gustavson say that "all that happens in self-deception . . . is that the person believes or forgets something in certain circumstances" and the

circumstances are such that the evidence does not warrant the belief in question (34-35). Jeffrey Foss writes that "Jones deceives himself that *p* just in case (i) Jones brings it about that *jBp* [Jones believes that *p*], and (ii) *jK~p* [Jones knows that not *p*]" (241).

I could continue this list, but it is already clear that there is no agreement among philosophers on what self-deception is. Fingarette thinks that self-deception concerns engagements in the world, while the rest of the mentioned philosophers talk about beliefs that *p* and not-*p*. Foss insists that self-deception requires two contradictory beliefs, while Audi, Canfield, Gustavson, Siegler and Paluch relate self-deception to the presence of one unwarranted belief. Foss' phrase that Jones brings about a belief suggests that Jones intentionally deceives himself, Paluch's mentioning of 'motive for discounting' and Fingarette's 'disavowing of the engagement' could suggest the intention on the part of the self-deceiver, but the rest of the interpretations do not imply such intentional deception at all. Audi thinks that self-deception requires an unconscious knowledge, while other philosophers do not mention either unconscious knowledge or beliefs.

One could think that at least some philosophers have defined self-deception incorrectly, and a good analysis of the concept would eliminate the multiplicity of interpretations. No doubt, philosophers do argue about the proper way to define self-deception. For example, Foss thinks that "Gustavson, Canfield, & co. [meaning Patrick Gardiner and Terence Penelhum]" have missed a very important aspect of self-deception -- duplicity (238). At the same time, despite the criticisms of one or another definition, there is still a great diversity among the interpretations that one can find in the literature on self-deception. And it seems that these interpretations exist side by side without undermining each other's validity. At least, some philosophers are ready to admit the validity of interpretations different from their own (See, for example, Siegler 475).

I think that the diversity of interpretations asks for an explanation. Even if the topic of my thesis does not compel me to explain this diversity, I cannot ignore it. In order to proceed to the question of deliberate self-deception, I must have some understanding of

what self-deception is. If I just accepted one definition that seemed more suitable for the topic of my thesis, I would not know how this definition is related to the others. I would not know whether different interpretations of what self-deception is are different definitions of the same phenomenon, or they are descriptions of different phenomena under one name, or descriptions of different usages of the word 'self-deception.' Since I am interested to know what other philosophers have said about the possibility of deliberate self-deception, I cannot just choose one, more convenient, interpretation, because there is no reason to presume that it will allow me to understand what other philosophers mean when they are talking about self-deception. Therefore, in this chapter I want to explain the great variety among the definitions of self-deception, and I think that by doing this I will be able to clarify for myself what self-deception is.

To find my way through this multitude of interpretations, I want to concentrate my attention on the *term* 'self-deception' and the meaning of this term in ordinary language. When I say that I will concentrate my attention on the meaning of the word in ordinary language, I do not intend to say that I will just describe the usage of the word. I want to use the meaning of ordinary language as a basis for comparison of different interpretations. I want to detect how close or how far from the ordinary meaning these interpretations are. My choice of the basis of comparison is not arbitrary; 'self-deception' is a word of ordinary language and was used before philosophers started to discuss this concept; I presume that philosophers' understanding of 'self-deception' has something to do with 'self-deception' of ordinary language, otherwise it would be hard to understand why philosophers use this word. I think that by revealing this correlation I will be able to answer the question, "What is self-deception?"

In Section 2.1, I concentrate my attention on the claims that self-deception must be understood as interpersonal deception carried out by a person on himself or herself. This method of defining 'self-deception' results in a paradoxical account of self-deception. I claim that this method of defining self-deception cannot provide one with the understanding

of the ordinary language term 'self-deception.' Nevertheless, I admit that the ordinary language term 'self-deception' has a paradoxical meaning. 'Self-deception' means making oneself believe what one knows is false. In Section 2.2, I describe the non-paradoxical meaning that the term 'self-deception' has in ordinary language. Before providing this non-paradoxical meaning, I have to reject two plausible versions of this meaning. In the first part of Section 2.2 "Unusual Ways of Deceiving Oneself," I show some occasions when a person could be described as deceiving oneself, but only if the expression 'to deceive oneself' is used in some particular sense that differs from the use of the term 'self-deception' in ordinary language. In the second part of Section 2.2 "'Deceiving oneself' and Unwarranted Belief," I analyse and ultimately reject one interpretation of the term 'self-deception' that describes self-deception as a discrepancy between some belief and evidence for this belief. I claim that this interpretation depicts the usage of the expression 'to deceive oneself' and not the meaning of the term 'self-deception.' In the third part of Section 2.2 "'Self-Deception' and Biased Believing," I provide one interpretation of the term 'self-deception' that seems to me a correct description of non-paradoxical meaning if the term in ordinary language. In Section 2.3, I try to explain the what is comment between the paradoxical meaning of 'self-deception' that the term has in ordinary language and the variety of definitions of 'self-deception' provided by philosophers.

2.1 Self-Deception and Other-Deception

A natural way to find out the meaning of the word 'self-deception' seems to be consulting a dictionary. Unfortunately, a standard English dictionary is not much of help to me. The OED for example defines self-deception as an act or state of deceiving oneself. This definition is circular and is not informative. It explains that self-deception can be an act or a state, but it does not tell me what kind of act or state self-deception is, whereas I am interested in knowing exactly the nature of this state or act.

At the same time, one could interpret this definition as a suggestion that self-deception is a particular case of deception, where 'deception' has to be understood on the model of interpersonal deception. This approach could work approximately like this: (a) the word 'self-knowledge' consists of two parts, 'self' and 'knowledge;' when I know what 'knowledge' means and what 'self' means, I can easily deduce that self-knowledge is just like knowledge only the subject of knowledge is specified -- the self; (b) the word 'self-deception' consists of two parts, 'self' and 'deception;' when I know the meaning of both of them I will know the meaning of 'self-deception.' To find out the necessary and sufficient conditions for ascribing the word 'deception,' I analyse, for example, a sentence 'John deceives Peter;' to know what self-deception is, I simply replace 'Peter' with 'himself.'

There are some philosophers who accept this way of understanding 'self-deception,' and one of them is Raphael Demos. Demos' article "Lying to Oneself" is the article that brings the *term* 'self-deception' into the sphere of philosophical interest. Even though the first philosopher who mentions self-deception is most likely Plato, before Demos 'self-deception' has not been among the concepts that inspire philosophers. Self-deception has been discussed among Christian moralists, for example, Samuel Johnson and Bishop Butler have articulated their perception of self-deception, and Daniel Dyke's book *The Mystery of Self-Deception*, which was written in the beginning of 17th century, is most likely the first book on self-deception. Nevertheless, neither of these moralists finds the concept of self-deception in any way puzzling. As ordinary users of language, they know when to apply the word and are not interested in spelling out its meaning. When Demos tried to define the concept of self-deception and analyse its implications, he quickly provoked a criticism of his definition, thus starting the discussion. On the whole, philosophers dismiss Demos' analysis of 'self-deception' as incorrect. At the same time, Demos' article on self-deception 'gave a tune' for the later discussion of self-deception, and some aspects of this discussion are hard to understand unless one knows what Demos did with 'self-deception.'

The problem of self-deception, as it is stated by Demos, resembles a puzzle meant

to sharpen one's mind. Demos begins his article "Lying to Oneself" with laying down the conditions of the intellectual exercise (588). First of all, one has to assume that words 'lying' and 'deceiving' have identical meaning. Demos recognizes that the meanings are not identical, but he asks the reader to ignore this fact. Secondly, one has to assume that the phrase "B lies to (deceives) C" means that the deceiver, or liar, *intends* to induce a mistaken belief in another person and *succeeds* in carrying out his intention. Moreover, the deceiver *knows* that what he tells another person is false. Demos acknowledges that one can deceive a person without intending to do so and that one can lie without causing anybody to believe one's lies; nevertheless, Demos deliberately disregards these aspects of deception. Only after one has accepted both conditions, is Demos ready to show the problem in which he is interested. He reformulates his description of interpersonal deception so that the act of deception is presented as occurring within one person. According to Demos, "self-deception exists . . . when a person lies to himself, that is to say, persuades himself to believe what he *knows* is not so" (588). Thus stated, self-deception seems to be impossible. For example, one can try to persuade oneself to believe that grasshoppers eat people, and most likely one will fail. To make things worse, Demos interprets his formulation as implying that the self-deceiver believes some proposition *p* and the negation of this proposition at the same time (588). He also declares that both beliefs are consciously held (592). Thus Demos has formulated what philosophers like to call 'the paradox of self-deception,' because it seems impossible to believe in a proposition that one knows to be false. Alfred Mele calls this paradox the 'static paradox,' which he distinguishes from the 'dynamic' paradox ("Recent" 1). The term 'dynamic paradox' is used to describe the apparent difficulty of making oneself believe something that is known to be false. The challenge in Demos' puzzle of self-deception is to find out how such intentional and paradoxical self-deception is possible.

A specific characteristic of Demos' treatment of the concept of self-deception is his indifference to other possible meanings of this concept. He is not interested in knowing

whether 'self-deception' in everyday language has the same meaning as 'self-deception' in his formulation. Describing the conditions that must be realized in order for us to call something self-deception, he chooses conditions similar to those of 'deception' in the interpersonal context, and he does not inquire whether it is possible to define 'self-deception' otherwise. In addition, Demos ignores other possible meanings of the word 'deception.' Since according to Demos the concept of self-deception is derived from the concept of deception, he attributes to 'self-deception' a very specific and narrow sense.

This lack of interest on the part of Demos does not make other interpretations of self-deception less real. Unfortunately, it is not difficult to overlook the variations. To illustrate how different meanings of 'self-deception' can be confused, I want to show an understanding of self-deception that is radically different from Demos' understanding and which Demos himself incorrectly equates with the one he presented in his article. I am referring to Plato's concept of self-deception. Demos claims that his and Plato's understanding of self-deception are the same (588). Plato mentions self-deception rather casually in the dialogue *Cratylus*, and he does not explicate what precisely he understands by it. Despite Plato's terseness, it is possible to tell the difference between his and Demos' understanding of 'self-deception.'

Plato's dialogue *Cratylus* discusses the question of how names relate to the things they name. At one place in the dialogue, the discussion of names is interrupted by the exchange of compliments about the wisdom of the interlocutors. Cratylus is so impressed by everything Socrates says that he suggests that some Muse resides in Socrates and speaks through him (428c). Socrates agrees and complains that he cannot trust his own wisdom and words he utters. In this context, Plato says that there is nothing worse than self-deception, because "the deceiver is always at home and always with you" (428d).

It is possible that Plato suggests the existence of some spiritual entity that resides somewhere inside a human being and can be truthful or deceptive. It seems to me, nevertheless, that the remedy against self-deception Plato indicates implies a more

interesting understanding of 'self-deception.' In order to avoid deceiving himself, Socrates has to examine all the claims he makes and retrace the course of his argument, or "steps" (428d). If in fact Socrates spoke for some Muse, or some other divine being, then the scrutiny of thoughts would be useless; Socrates could not influence his *alter ego* even if he wanted. Moreover, the multiplicity of personality would not explain why Socrates describes the deception as deception of *oneself*. I am suggesting that Socrates speaks about his thoughts as if they were imposed upon him because he does not understand how the thinking process works and how thoughts are generated. Nevertheless, he is not alienated from his thoughts, because he knows how to control them. He controls his thoughts by analyzing them in retrospect.

It seems to me that for Plato, self-deception characterises the process of reasoning or, more precisely, erroneous reasoning. Self-deception is a mistake that is caused not so much by wrong information as in cases of deception, but rather by inaccurate thinking. Self-deception is worse than deception, because erroneous thinking affects one constantly, while deceivers are not always around. Also, it is very hard to notice the failure of one's reasoning, and even Socrates cannot be sure that his wisdom is not deceptive. The only way he can control this kind of deception is to review and analyse his train of thoughts, and to do that often. The analysis needs not to be done by the thinker alone. Just before Socrates mentions self-deception, he encourages Cratylus to criticize everything Socrates says (428b). Discussion is one way to detect faulty thinking and, therefore, discussion undermines the possibility of self-deception.

If my interpretation of Plato is admissible, it is hard to see how one can equate Plato's understanding of self-deception with the interpretation of 'self-deception' that Demos presents in his article. Demos declares that the self-deceiver *intends* to make himself believe something that is false, while *knowing* that what he wants to believe is false. Plato's self-deceiver does not need to have either the intention to deceive himself, or the knowledge of the falsity of his beliefs. For Plato, one must work hard to notice the

falsity of one's beliefs.

It seems that Demos is mistaken when he claims that his understanding of 'self-deception' is similar to Plato's. It is not similar. Now one can ask why it is not. Are Plato and Demos using different meanings of the same word? Does Plato misuse the term? What he calls 'deception of oneself' seems to fit better under the name 'faulty reasoning.' Does Demos misuse the term?

There are two main objections against Demos' interpretation of 'self-deception.' First, philosophers argue that the word 'deception' need not necessarily imply either that deception is carried about intentionally or that the deceiver knows that the proposition he wants others to believe is false. Mele claims that sometimes people use word 'deceive' in cases when somebody unintentionally causes another person to believe some proposition that is false (*Irrationality* 123). Bas Van Fraassen argues that the deceiver can be ignorant about the truth of the proposition which he or she wishes to deceive others into believing. For example, if Peter does not know whether some bridge is safe, but he wants John to believe that the bridge is safe, Peter could deceive John by persuading him that the bridge is safe (124). Brian McLaughlin claims that the deceiver can even believe in the truth of the proposition about which he or she wants other people to be deceived. For example, evidence appears to prove that Dick is guilty of some wrong-doing; Tom believes that Dick is innocent and by lying persuades Harry to believe in the innocence of Dick (35). Since Demos derives his definition of 'self-deception' from the definition of 'deception,' changes in the latter can cause changes in the former.

The problems with the definition of 'deception' do not undermine the main principle of defining 'self-deception' on the basis of how the word 'deception' is understood. This principle seems very sensible. The connection between the concept of self-deception and the concept of deception looks evident: the term 'self-deception' includes the word 'deception.' If I know what is 'deception' is, I can explain what 'self-deception' is. At first, I analyse the meaning of the concept of deception in the interpersonal context; next, I

describe a pattern of deception in a case when deception is not directed towards another person, but towards oneself. This approach demonstrates the reason for the presence of the word 'deception' in the concept of self-deception, while for example in the case of Plato, it is not clear why one would talk about deception at all.

Despite the appealing simplicity of Demos' approach, not all philosophers like his way of defining 'self-deception.' Several of them have argued that 'self-deception' cannot be analysed in the same terms as 'deception.' This is the second and the most important objection against Demos' definition. It is more important than the first one, because if it is true that the meaning of 'self-deception' cannot be obtained by analysing the meaning of 'deception,' then it is not important for the defining of 'self-deception' how the word 'deception' is interpreted.

The first philosophers who argued against Demos' method of defining 'self-deception' are Canfield and Gustavson. They claim that any explanation of 'self-deception' using the concepts of interpersonal deception requires the presupposition that 'self-deception' can be properly understood only in terms of interpersonal deception, or other-deception (32). Canfield and Gustavson defy this presupposition and show that in other-deception and self-deception 'deception' can mean different things.

The argument by Canfield and Gustavson is directed against the method of defining that was used by Demos, namely, to explain terms like 'self-deception' by explicating the part of the term that comes after 'self-'. If this method were correct, then 'self-command' should be understood as being similar to 'command,' or 'other-command,' specifying that the commander was identical with the person that received the command (33). Or, using Mike V. Martin's example, 'teaching oneself' should be understood in terms of teaching others (19). Canfield and Gustavson claim that 'self-command' cannot be understood in terms of other-command. To justify their claim, they first of all disclose the assertions that are implied by the notion of command. Then, they apply these assertions to the notion of self-command. They believe that the result demonstrates that this juxtaposition of 'other-

command' and 'self-command' is inappropriate.

Canfield and Gustavson consider one instance of 'other-command' that they formulate as 'Jones makes Smith do E.' The formulation is strange because to say that 'Jones commands Smith jump' is not the same as saying 'Jones makes Smith to jump.' The latter implies that Smith in fact jumps, while the former need not imply that: Jones commands, but Smith ignores him. Canfield and Gustavson seem to be talking about a successful command. It is possible that Canfield and Gustavson wanted to emphasize the similarity between 'command' and 'deception' which, according to the standard interpretation, implies that the deceiver succeeds in deceiving the other person.

According to Canfield and Gustavson, the assertions that are implied by the statement 'Jones makes Smith do E' are: (a) Jones intends to make Smith do E; (b) Jones asks (commands, tells,etc.) Smith to do E; (c) Smith takes Jones' request (command, etc.) as a request to do E; (d) Smith complies with (obeys, etc.) Jones' request to do E (33). If one wants to present an instance of a successful self-command and to interpret it as similar to the case of other-command, the sentence 'Jones makes himself study all night' must be interpreted as implying that "Jones intends to make himself study all night, Jones asks (commands, etc.) himself to study all night, Jones takes his own request as a request to study all night, and Jones complies with his own request to study all night" (33-34). It is clear that on this interpretation of 'self-command' the part that corresponds to (c) is redundant. Jones does not have to interpret his own requests and orders in addition to saying them. I also would deny that the word 'complies' can be used to describe the connection between a command Jones utters and the action that follows. Therefore, Canfield and Gustavson suggest that 'self-command' cannot be understood in terms of other-command. Correspondingly, the whole method of explaining the concepts that have the form 'self-*x*' by first explaining the part that follows 'self-' is undermined, because Canfield and Gustavson have shown that there is one instance where this method does not work.

Martin comes up with another example where, according to him, the method used by Demos cannot provide the proper understanding of the term. He uses the example of 'teaching' and 'teaching oneself.' One small problem with this example is that there is no such a term as 'self-teaching' in ordinary language. Neither 'self-taught' nor 'teaching oneself' are strictly parallel to the term 'self-deception.' And as I will show later, 'self-deception' is not always replaceable by 'deceiving oneself.' At the same time, I think that this problem does not really undermine Martin's idea, because there is a term that is very similar to Martin's 'teaching oneself,' and the term is 'self-instruction.' This term also has the related forms 'self-instructed' and 'instructing oneself' which make it similar to Martin's 'teaching oneself' and Demos' 'self-deception.'[1]

The examples of 'teaching' and 'instructing' are interesting because of one important resemblance with 'deception.' If the meaning of 'teaching oneself' or 'self-instruction' is derived from the meaning of 'teaching' and 'instructing,' it seems that the first two will be as paradoxical as the concept of self-deception. The concept of teaching usually implies that one person knows something that another person does not (Martin 19). The same is true about 'instruction.' If one knows everything that the instructor is telling one, one hardly would call this process an instruction. So if one wants to consider 'self-instruction' as the correlate of 'instruction,' one has to assume that persons who instruct themselves both know and do not know how to perform some action, know and do not know the content of instructions. If in the case of 'self-deception' the paradoxical meaning seemed probable, in the case of 'self-instruction,' the interpretation obtained by juxtaposing 'instruction' and 'self-instruction' clearly gives wrong results. When I am saying that I am instructing myself, I am saying that I am learning to do, or I am doing, something without an instructor, or without knowing beforehand the instructions, and I am not suggesting that I

[1] The example of 'self-instruction' is used also by William Ruddick in his article "Social Self-Deception" (384-385).

somehow simultaneously know and do not know these instructions.

Now one could ask why 'self-instruction' is called 'instruction,' if in fact it is nothing more than learning. It is very hard to answer the question why such a term has taken roots in the language, but I can indicate some correlations between 'instruction' and 'self-instruction.' First of all, self-instruction is a process of doing something without instructions while usually one would not do this without them. For example, I can try to learn languages on my own, while usually one would ask somebody for instruction on how to do this. Or I can try to build a house or play piano without previous knowledge of how to do this. On all of these occasions I can say that I am self-instructed, because nobody else has instructed me. Secondly, I would use the word 'self-instruction' to describe a situation when I am using instructions that are prepared by some person who is absent. For example, I am using a book *How to Build Houses*. The instructions are given by the person who wrote the book, but there is nobody who will inform me of these instructions except myself and, therefore, I can say that I am instructing myself.

It seems that there are reasons for using the word 'instruction' to describe actions like learning languages without a tutor. The problem is that one cannot know in advance how the word 'instruction' must be used in order to explain 'self-instruction.' It is not the meaning of 'instruction' that explains the meaning of 'self-instruction,' but the meaning of 'self-instruction' explains which aspect of the word 'instruction' one has in mind when one speaks of 'self-instruction.'

The same is true about the way one understands 'a successful self-command.' When Jones makes himself study all night, he does not need to command himself to study all night and take his own command as a command to study all night, as the meaning of 'make somebody to do something' implies. At the same time, one can articulate one's intention to study all night and do it in the form that resembles an interpersonal command. There will be resemblance between the articulation of an intention and the interpersonal command, but one has to know how 'self-command,' or 'make myself do something,' is used in order to

detect what the resemblance is.

Now I can explain why Plato calls erroneous reasoning 'self-deception.' He does not derive the meaning of deception by analyzing interpersonal deception. The erroneous reasoning can be called 'self-deception' because a person that errs can be viewed as being misled. When somebody deceives me, he misleads me. If I am misled by my failure to reason properly, I can say that I deceived myself.

When Demos defines 'self-deception,' he determines at first the meaning of 'deception' and then, depending on the meaning of 'deception,' determines the meaning of 'self-deception.' This method cannot guarantee that one will be able to understand what is meant by 'self-deception' in ordinary language, or what Plato means by 'self-deception.' In both cases, the meaning of 'deception' can be used in a quite different way than it was used by Demos. Since Demos' method of clarifying the meaning of a concept cannot guarantee a reliable interpretation of the concept, the definition of 'self-deception' that is formulated in terms of other-deception cannot be binding for anybody who is asking how 'self-deception' must be understood. There must be some other way to determine the meaning of the concept 'self-deception,' and I think that the other way is to describe how the concept is used in ordinary language.

Demos' definition of 'self-deception' required for a self-deceiver to know simultaneously that some proposition *p* is true and to believe that *p* is false, and it also required that the self-deceiver somehow intentionally *make himself or herself* believe what he or she knows is false. I showed that the method Demos used to define 'self-deception' cannot provide me with the meaning that this term has in ordinary language. One could think that if Demos' interpretation of 'self-deception' is not binding upon anybody studying self-deception, one is free from the hardest part of the problem: paradoxes of self-deception. Demos' idea of defining 'self-deception' on the model of 'deception' seemed to create both the static and dynamic paradoxes of self-deception; since Demos' method of defining turned out to be unreliable, one might assume that one can just reject the

paradoxes as a result of faulty thinking.

The strange thing is that the analysis of the everyday meaning of 'self-deception' seems to throw me back where I started. Demos' definition presents quite precisely one of the meanings that 'self-deception' has in ordinary language. Let me look at an example of self-deception: some boy is cruel to animals; his mother has seen some occasions when he killed seven bumble-bees, and other people have reported to her similar episodes in her son's life; nevertheless, she denies that her son is cruel to animals, and it seems that she really believes what she says. This is a situation when one could ascribe to the mother self-deception. Asked what one means by this claim, one could say that the mother knows that her son is cruel to animals (after all, she saw him being so), but she intentionally ignores the evidence and makes herself believe that he is a good boy. At this point in my analysis, it is not important whether the mother really knows that he is cruel or whether she intentionally ignores this knowledge, I just want to clarify what people would mean by saying that the mother is self-deceived. I think that 'self-deception' of ordinary language has the meaning I described, and evidently, this meaning is similar to the meaning that Demos presented in his article.

Initially, the fact that Demos' interpretation of 'self-deception' coincides with one of the meanings that 'self-deception' has in ordinary language could look a little bit embarrassing. Twenty pages of my thesis are spent to prove that Demos' method is inadequate just to find out that Demos' interpretation is a quite conventional interpretation of self-deception. Nevertheless, I dare to claim that these pages are not just a collection of vanities. There are three conclusions that this analysis has helped me to reach.

The First Conclusion. The fact that the word 'deception' is usually used to describe an act by which a person deceives some other person does not have to imply that 'self-deception' designates the same act only carried out on oneself. So I would object, for example, to Frederick F. Schmitt's assertion, "If there is genuine self-deception, properly so called, it must consist of deceiving oneself into believing some proposition" (189).

Consistent with his claims, Schmitt continues with ascribing to 'self-deception' the conditions of interpersonal deception. My main objection against his claim concerns the usage of the words 'genuine' and 'properly.' As far as I can see, everything that in ordinary language bears the name 'self-deception' is genuine self-deception. Even if the name is used to denote something that is not like internal deception, I cannot see the reason for claiming that it is not genuine self-deception.

I also doubt that one can say that the absence of an act of deception somehow makes the name 'self-deception' inappropriate. 'Self-deception' is a word of ordinary language and, as far as ordinary language is concerned, to question the choice of words for designation of one or another phenomenon seems to me rather fruitless enterprise. Are butterflies named properly? Do genuine butterflies, properly so called, have anything to do with butter? Should one analyse the words 'butter' and 'a fly' to know what 'butterfly' should properly mean? I think that these questions may be asked when one tries to invent a new name for something, but I cannot see any reason to ask them about a word of ordinary language.

Just to give the reader a feeling of how confusing for philosophers the word 'self-deception' has turned out to be, I want to mention one more difficulty that concerns a 'proper' understanding of 'self-deception.' Several philosophers have presented Mary Haight as assuming that 'self-deception' properly understood has to be interpreted as interpersonal deception within one person (for example, Martin 18 and Mele "Recent" 2). What Haight says is that "if to deceive oneself is really to deceive, a definition of 'A deceives B' should fit some cases where B and A are the same, and these should be the cases that in fact we call 'self-deception'" (8). 'Really to deceive' may sound like 'self-deception properly understood,' but in fact Haight is just saying that if 'self-deception' is understood as interpersonal deception within one person (understood literally), then there must be phenomena that correspond to the definition and these phenomena must be called in ordinary language 'self-deception.' As far as I can see, she is not claiming that 'A

deceives A' is the proper understanding of 'self-deception.' She just wants to clarify whether 'self-deception' *could* mean this. She concludes that it could not, that there cannot be deception within one person and that the term 'self-deception' must be understood as a metaphor or a figure of speech (23,52). And she does not claim that the metaphorical expressions are 'improper' expressions or that any understanding of 'self-deception' which does not depict self-deception as deception within one person is improper.[2]

The Second Conclusion. Now it is clear that the reflections on what are the sufficient and necessary conditions for ascribing to somebody 'deception' are interesting in themselves but not very helpful for understanding the sufficient and necessary conditions for ascribing 'self-deception.' The understanding of the word 'deception' can be helpful for understanding the reasons for calling self-deception 'deception,' like understanding of the words 'grass' and 'to hop' can be helpful for understanding why certain insects are called 'grasshoppers.' Nevertheless, even if one or several interpretations of 'deception' could be used to explain the meaning of the word 'self-deception,' one would not be able to tell which ones without knowing in advance what 'self-deception' means. So, I can enjoy Van Fraassen's, McLaughlin's, Schmitt's, Anette Barnes' or Stanley Paluch's thoughts on what are the necessary conditions for something to be deception, or what are the correlations between 'self-deception' and 'deception,' but I cannot use these ideas by themselves to determine the meaning of the term 'self-deception.'

The Third Conclusion. Even if the meaning of 'self-deception' cannot be derived from the analysis of 'deception,' the word 'self-deception' still can have a meaning that is apparently paradoxical. This conclusion can be rephrased in the form of an instruction for those who try to find their way in the writings on self-deception: 'Do not trust anybody who claims that the paradox of self-deception stems from the efforts to derive the concept of

---

[2] Similarly, Martin misunderstands Kipp's interpretation of 'self-deception' (Martin 18, Kipp 261, 279).

self-deception from the concept of deception.' Unfortunately, many philosophers claim or imply this origin of the paradoxes. For example, Mele writes, "In both cases [in cases of the static and dynamic paradoxes], paradox is generated by the application of certain common assumptions about interpersonal deception to the *intra*personal variety" ("Recent" 1). David Pears, "How can anyone persuade himself that *p* and yet all the time maintain his original belief that *not-p*, as the word 'deception' seems to require?" ("The Goals" 59). Martin, "The air of paradox arises when we try to understand self-deception by modelling it strictly after interpersonal deception (that is, the deception of one person by another)" (13). If they were right, it would be easy to get rid of the paradox. One could simply claim that the paradox can be ignored because the meaning of 'self-deception' does not depend on the meaning of 'deception,' and describe the meanings of 'self-deception' that are not paradoxical. Unfortunately, one cannot ignore the paradox, because it does not depend on the meaning of 'deception.'

2.2 Non-Paradoxical Meaning of 'Self-Deception'

Are there any meanings of the term 'self-deception' without the paradoxical one? As far as I can see, the only way to answer this question is to describe the meanings that the word has in ordinary language. Such description can be problematic. It is hard to know when one has described all existent meanings; the only criterion is one's knowledge of language. No philosopher has attempted to present an exhaustive description of the meanings that 'self-deception' has in ordinary language. I will not attempt to do it either, but I will present and examine some explications of the meaning that have been discussed by philosophers, and in the end of this chapter I will describe a meaning of 'self-deception' that the term has in ordinary language and that is not paradoxical.

2.2.1 Unusual Ways of Deceiving Oneself

Several philosophers have provided examples of persons who deceive themselves but still cannot be considered self-deceivers. For example, it is reasonable to say that a military camouflage expert has deceived himself when he has disguised the field gun so well that he cannot recognize from distance where exactly the gun is hidden (Champlin "Deceit" 57). It is reasonable to say that a cocaine dealer has deceived himself when he, by submitting himself to a seance of hypnosis, makes himself believe that his supplier is Ronald Reagen (Silver 216). As the authors of the examples have recognized, neither of the cases represents self-deception, and they conclude that the deception of oneself is not always what is called 'self-deception.'

Although I agree that the examples mentioned above are not examples of self-deception, I must make a brief comment on why these examples are not examples of self-deception. Maury Silver, John Sabini and Maria Miceli have noted that their example of the dealer is not "an example of what people call 'self-deception," but they also immediately add that the reason why it is not the right example is because the goal of the deceiver is not to manipulate his feelings but something else (Silver 216). According to them, the goal to manipulate feeling is essential for something to be self-deception. Similarly, T. S. Champlin claims that the example of the camouflage expert is not an example of self-deception, because the aspect of dishonesty with oneself and moral shortcoming is missing ("Deceit" 57). The problem with the claims about what is missing from these examples is that the conditions which the philosophers claim are absent are not necessary for using the word 'self-deception.' And I will show later why they are not. Meanwhile, I want to say that ignorance about the necessary conditions for ascribing to someone 'self-deception' cannot prevent one from dismissing the examples of the drug dealer and the camouflage expert. I think that anybody who knows English knows that the word 'self-deception' is not used in ordinary language to describe such cases. That is simply not the way 'self-deception' is used, and our knowledge of that is enough for making

a distinction between 'self-deception' and 'deceiving oneself' as it is used in the examples of the camouflage expert and the drug dealer.

2.2.2 'Deceiving Oneself' and Unwarranted Belief

I have already described the paradoxical meaning of 'self-deception' that is ascribed to the word in ordinary language and that has proved to be a problem for anybody who tries to interpret it. At least sometimes, 'self-deception' means that one makes oneself believe what one knows is false. The paradoxical nature of this formulation has caused philosophers to look for alternative interpretations. Several philosophers have discussed the possibility that 'self-deception' could mean that a person has not noticed something obvious, at least something that seems obvious to a person who ascribes self-deception to somebody.

Canfield and Gustavson emphasize the ignorance of the obvious and consider it being the basic characteristic of the phenomenon that is called 'self-deception.' They claim that "when Jones deceives himself about P, he believes P in belief-adverse circumstances, or he forgets P when, ordinarily, one would remember P" (36).

The notion of belief-adverse circumstances is a little bit ambiguous. Patrick Gardiner, for instance, thinks that Canfield and Gustavson's definition can be interpreted as claiming that some person believes *p* while disinclined to believe *p* because *p* seems to have unpleasant implications (Gardiner 229). For example, John can realize that his belief that smoking damages lungs implies that he should quit smoking, and while John is reluctant to quit smoking and challenges any proof that smoking damages lungs, he still believes that it does. One could say that John believes 'smoking is harmful' in a belief-adverse circumstance, which in this case is the fact that John tries to defy his belief. Despite the plausibility of Gardiner's interpretation, one cannot accept it, because Canfield and Gustavson are quite clear about what they mean by 'belief-adverse circumstances.'

They mean "circumstances such that the evidence Jones has does not warrant belief in P" (34). To say that 'belief-adverse circumstances' means evidence that does not warrant belief in *p* is not the same as saying that 'belief-adverse circumstances' means disinclination to believe that *p*. Hence, Gardiner's interpretation seems to be inadequate.

What causes the misunderstanding is Canfield and Gustavson's suggestion that self-deception must be treated as a special case of self-command. According to them, the sufficient condition for ascribing 'self-command' is that somebody does something in the face of certain obstacles, for example, one studies despite the disinclination to study (34). Most likely, Gardiner thinks that Canfield and Gustavson are saying that just as one can make oneself study while being disinclined to study, so one can believe something while being disinclined to believe it. Nevertheless, Canfield and Gustavson do not mention this interpretation of 'belief-adverse circumstances,' neither do they mention how their interpretation follows from the supposed similarity between 'self-deception' and 'making oneself to do something.' If they would claim simply that they will presuppose that 'self-deception' means doing something in the face of certain obstacles, one could accept it as a heuristic device and examine whether this supposition is adequately presents the meaning of 'self-deception' in ordinary language. Nevertheless, Canfield and Gustavson claim that they will treat self-deception as "a special case of self-command" (34). It seems to me that if self-deception is a special case of self-command, one should understand what is the connection between self-command and self-deception. One should understand why believing in something that is not warranted by evidence, that is to say, believing in belief-adverse circumstances, should be considered as a special case of self-command

For example, they must explain what they understand by the words 'do' and 'believe.' I can say that in some sense to believe in belief-adverse circumstances is to do something. John says: "I believe in Santa Claus." Mary is surprised: "Do you?" Nevertheless, believing is certainly not an action. I do not do believing in the same sense I do catching of a water measurer. Meanwhile, Canfield and Gustavson's interpretation of

self-command implies an action: I am doing something in the face of certain obstacles, for example, studying despite tiredness (34). If 'doing in the face of certain obstacles' is meant to imply an action and believing in belief-adverse circumstances is 'a special case of self-command,' Canfield and Gustavson must say that believing in the face of evidence is an action too. I can imagine only one sense in which 'believing' designates action, namely, in case when 'believing' is a shorter way of saying 'making oneself to believe something.' Hence, to say 'I believe in belief-adverse circumstances' is to say that 'I make myself believe in belief-adverse circumstances.' I think that this interpretation is the same paradoxical account of 'self-deception' that Canfield and Gustavson wants to refute. At the same time, if Canfield and Gustavson want to ascribe to the word 'doing' a very broad sense, they will end up with rather dubious examples of self-command. Mary who *looks young* despite her age also *does* something in the face of certain obstacles. Nevertheless, I am reluctant to call this example an example of self-command.

Because of certain shortcomings of Canfield and Gustavson's analysis, I must agree with Herbert Fingarette that their definition of 'self-deception' must be considered on its own merits, ignoring the way they obtain this definition (Fingarette 22). At first, it seems strange to suggest that one should analyse a definition ignoring the way it is acquired. Nevertheless, one must remember Demos whose definition of 'self-deception' adequately presented the meaning of 'self-deception' despite the flaws in the method that Demos used to acquire this definition. As in the case of Demos' definition, Canfield and Gustavson's definition seems to capture one of the meanings that the concept 'self-deception' has.

It seems that at least sometimes people mean by 'self-deception' nothing more than ignorance of the obvious. For example, Jacques Derrida says in an interview, "In all the other disciplines [economics, sociology, the natural sciences, literature] you [Richard Kearney] mention, there is philosophy. To say to oneself that one is going to study something that is *not* philosophy is to deceive oneself" (Kearney 114). It is hard to see what Derrida would mean by this phrase except that the imaginary student errs in his or her

thinking while it is quite obvious, according to Derrida, that any discipline has its share of philosophy.

There are several philosophers who have provided similar examples. M. J. Scott-Taggart, for example, says that "in everyday practice we frequently do use the falsity of someone's belief as a sufficient basis for a charge of self deceit. One frequently hears statements such as: "He is deceiving himself if he thinks I am going to visit her because he asked me to, for I am not" (11). Frederick Siegler tells the story about Brown whose wife is unfaithful. Brown confides to his friend, White, that it seems that his wife's friendship with her friend can lead her to infidelity, and White says that Brown is deceiving himself if he thinks that his wife is still faithful. According to Siegler, "White is telling Brown that he has an erroneous belief, and he is implying that it is unreasonable for Brown to have such a belief" (473). Similarly one can explain phrases like 'I am deceiving myself if I think I will win the race' or 'I am deceiving myself if I think I will go to China this summer' (474). It seems that Scott-Taggart's and Siegler's examples have shown that there is some truth in Canfield and Gustavson's definition.

Nevertheless, Canfield and Gustavson's definition of self-deception has provoked rather severe criticism. Its critics argue that the condition of belief in belief-adverse circumstances is not sufficient for defining self-deception. According to Penelhum and Gardiner, self-deception defined as believing in belief-adverse circumstances is not distinguishable from intellectual indecision, ignorance or stupidity (Penelhum 88), and error, confusion, ignorance or foolishness (Gardiner 231). I must agree that they are right in saying that Canfield and Gustavson's definition does not provide sufficient conditions for ascribing 'self-deception' to someone, and the examples that Gardiner and Penelhum provide in a way indicate several problems with this definition, but I also must say that their criticism has some flaws which for the sake of clarity should be mentioned.

First of all, one has to notice that there is certain vacillation going on between two different modes of talking about self-deception. I can talk about the term 'self-deception'

and some phenomenon that is called 'self-deception.' Canfield and Gustavson refer to both self-deception and 'self-deception,' deception and 'deception,' self-command and 'self-command.' When they come to define self-deception, they seem to talk about the phenomenon of self-deception and not the concept. For example, they say, "All that happens in self-deception . . . is that the person believes or forgets something in certain circumstances" (35). And Gardiner suggests that if that is all that happens then one cannot distinguish self-deception from, for example, ignorance. In the case of ignorance of certain evidence, nothing really happens except that a person believes in some proposition that is unwarranted. At the same time, he remarks that maybe there is no clear distinction between foolishness and self-deception, because, "it is possible to cite instances where saying of a person that he has deceived himself about a particular matter seems to come down to asserting no more than that his judgement was mistaken and that he should have known better" (231). Here Gardiner refers to Siegler's examples of the usage of the expression 'to deceive oneself' (Siegler 473-474). So it seems that Gardiner has to accept Canfield and Gustavson's definition after all. He solves the puzzle by announcing that "such uses [the ones mentioned by Siegler] appear to be peripheral and not to reflect the cardinal features of the concept as normally understood" (231). As one can notice, he has made a slip from talking about what happens in self-deception to what 'self-deception' means. At first he talks about what 'really happens' in cases of self-deception; later he makes claims about what people say and assert when they use the word 'self-deception.' I think that Gardiner has not noticed the ambiguity of Canfield and Gustavson's interpretation because of the similarity between their claims about self-deception and Siegler's claims about the usage of the word 'self-deception.' In fact, it is rather possible that they are claiming the same thing, but the ambiguity of the way philosophers express their ideas does not allow one to be sure.

The differences between the two modes are important. Not everything that can be said about a phenomenon can be said about the term that is ascribed to it. In order to illustrate what I mean, I will use an example about the rising of the sun. I can say about the

sun that it rises. By 'rising' I mean that the sun goes up. Or if I am more sophisticated I would say that the distance between the horizon and the sun increases. If I want to describe the phenomenon that is called 'the rising of the sun' I would usually say that what happens is that the earth rotates and because of this rotation my position with regard to the sun changes. Using Canfield and Gustavson's phrase, all that happens when the sun rises is that the earth rotates and my position with regard to the sun changes. Can I conclude that when I say 'the sun rises' I mean to say that the earth rotates and my position with regard to the sun changes. I think I cannot. First, my phrase about the sun's rising does not analytically imply the rotation of the earth, and the fact that people some time ago did not know that the earth rotates and, nevertheless, used the phrase about rising sun should prove this claim; secondly, the meaning of the phrase in ordinary language does not suggest the rotation of the sun. So it seems to me that the distinction between the phenomenon (all that happens) and the meaning of the term must be made. Of course, the ignoring of this distinction will not always cause misunderstandings. It does not seem to matter whether I define the word 'grasshopper' or describe the particular insect. Meanwhile, I think that in the case of 'self-deception' the distinction between the meaning of the concept and the phenomenon that this concept is ascribed to must be made.

I agree that self-deception as it is defined by Canfield and Gustavson cannot be distinguished from the case of a person who is ignorant about evidence: he or she believes in some proposition while the belief in the proposition is in fact unwarranted. Or somebody can be confused and not understand the evidence; even such a state of confusion would not be distinguishable from self-deception, if self-deception is just a belief in belief-adverse circumstances. Meanwhile, if one looks at the concepts of 'stupidity,' 'confusion,' 'ignorance,' or 'error,' none of them can be defined as 'believing in spite of evidence that does not warrant the belief.' So, one cannot substitute the sentence 'John deceives himself' with sentences like 'John is stupid' or 'John is fool.' By 'stupidity' and 'foolishness' one usually means certain personal characteristics that display themselves in beliefs, reasoning

or actions. Certainly, Derrida could say that those who think that they will study sociology but in so doing will not study philosophy are stupid, meaning that a person with minimal capacities of thinking should have arrived at such a thought. Maybe Derrida could have said that, but he does not. He says that philosophy is incorporated in sociology and natural sciences and some persons do not see that.

Nevertheless, I think that the examples provided by Gardiner and Penelhum are interesting, because it seems that when people say that one has deceived oneself they imply something more than the fact that one has an unwarranted belief. For example, they are not implying that one has unwarranted belief *because of ignorance*. Penelhum claimed that the self-deceiver must know the evidence because otherwise the state of self-deception would be indistinguishable from ignorance (88). And it seems that 'deceiving oneself' usually implies that one knows the evidence. I very well know why I will not go to China, and I know why my chances of winning the race are slim. By saying that one possesses evidence I do not mean to imply that one who deceives oneself necessary realizes the truth to which the evidence is pointing. There is no extra research necessary for somebody to realize that philosophy is part of every discipline, but the particular person might not realize that what he or she knows about natural sciences or sociology is evidence for the presence of philosophy in these disciplines.

Penelhum's objection, which is also presented by Gardiner (Gardiner 231), cannot be an objection against Canfield and Gustavson's interpretation of self-deception. They claim that Jones deceives himself when he believes in some proposition in belief-adverse circumstances, and the belief-adverse circumstances are such that the evidence *that Jones has* does not warrant the belief (34). 'The evidence that Jones has' seems to imply that Jones is aware of the evidence. Meanwhile, Penelhum's criticism can be applied neither to Siegler's nor Scott-Taggart's interpretation of what it means to say that one has deceived oneself. They both claim that phrases like 'John deceives himself' mean nothing more than the fact that John has unwarranted belief. Neither have mentioned that 'to deceive oneself'

implies knowledge of evidence, and I think they should have.

Like ignorance, stupidity does not seem to fit people that deceive themselves. I doubt that somebody would say that, for example, John is deceiving himself if the person thinks that John is stupid, foolish or mentally ill. If Siegler's interpretation seems to exclude the possibility that John is naive or foolish, one cannot say the same about Scott-Taggart's or Canfield and Gustavson's interpretations. Siegler has noticed that when one says about oneself 'I am deceiving myself' or one claims that somebody else is deceiving himself or herself, then one usually implies that "I *should* have known better, but I did not" or "he *ought* to know (have known) better" (474). I know that there is no chance to win the race; nevertheless, I believe that I will win, while I should have abandoned this belief because it is unwarranted. Brown knows that his wife more and more often is going on business trips together with her friend; nevertheless, he thinks that she is faithful to him while he should have realized that she is not. I think that here 'should' and 'ought' mean that the person who ascribes deception of himself or herself to somebody presumes that the person in question is capable of having arrived at the appropriate conclusion. White, who ascribes to Brown deception of himself, expects something from Brown. White thinks that Brown is capable of coming to the conclusion that his wife is unfaithful but he does not. Persons who believe that they will study the natural sciences without studying philosophy are capable of coming to the right conclusion but they do not. Usually I believe something only when my belief seems warranted, but this time I believe something that is not warranted. If saying that one deceives oneself necessarily implies that one is capable of either believing or not believing in belief-adverse circumstances, then I can explain why 'deception of oneself' is not ascribed to mentally ill, stupid, or naive persons, and why it is not ascribed to children or persons that are confused. One does not expect of children, stupid or confused persons that they will realize the evidence which is against their beliefs.

Jeffrey Foss has provided one more objection against Canfield and Gustavson's analysis of self-deception. He criticizes Canfield and Gustavson for allowing the possibil-

ity that a self-deceiver is right (238). It is possible to imagine a situation when one believes in something despite the evidence against one's belief and the belief turns out to be correct. Of course, if it turns out that Brown's wife, despite her many business trips in company of the friend, is still faithful to Brown, one would not correctly say that Brown has deceived himself.

In order to justify the claim that 'belief in belief-adverse circumstances' is not an acceptable definition for 'self-deception,' Foss shows that a belief "in the face of adverse evidence" can be true (238). If such a belief can be true, one cannot say that the person who has such a belief is deceiving himself or herself. To prove his claim, Foss uses an example of Smith and Jones who have fallen overboard quite far from the shore. Both Smith and Jones know on the basis of their experience that they are not good swimmers and both somehow believe that they will reach the shore. Smith succeeds, but Jones does not. Both believed in belief-adverse circumstances and, therefore, conform to the requirements of self-deception as they are presented by Canfield and Gustavson. Nevertheless, Smith cannot be self-deceived because his belief was correct. So it seems that Canfield and Gustavson's conditions of self-deception are not sufficient.

As Foss has recognized, one way to avoid his objection is to announce that in order for one to be self-deceived the belief that one holds in belief-adverse circumstances must be false. Foss claims that such a condition would be introduced *ad hoc* just to save the definition. According to Foss, the real problem with this definition is that it presents self-deception only as discrepancy between a belief and evidence. Foss claims that Canfield and Gustavson have forgotten a necessary aspect of self-deception, namely, the duplicity "with its implications of duality and deceit" (238).

I think that such a condition would not be an *ad hoc* condition. When somebody applies the concept of self-deception to person whose belief is correct, I would say that the concept is simply used incorrectly. For example, I can imagine that somebody named Wilson watches from the boat Smith and Jones struggling to get to the shore and says that if

they believe that they will make it they are deceiving themselves. I think that what Wilson wants to say is that for him it seems obvious that Smith and Jones will not reach the shore and they are mistaken when they believe that they will. Since Smith in fact reaches the shore, Wilson is mistaken in evaluating Smith capacities and his words about Smith deceiving himself were uttered mistakenly. Similarly, when I say that I know that it will rain tomorrow, but it does not rain, I have misused the word 'know' for one or another reason. I doubt that the absence of rain somehow makes *ad hoc* the requirement of truth for a correct ascription of knowledge.

One might be tempted to conclude that in ordinary language to ascribe self-deception to someone is to say that (1) he or she believes in some proposition that is unwarranted by evidence, (2) the person knows the evidence, (3) he or she is capable to adjust this belief to evidence, and (4) the proposition in which the person believes is false. Nevertheless, I want to claim that these conditions are not conditions for ascribing to someone the term 'self-deception.' I think these four conditions that seem to depict the meaning of the term 'self-deception' in fact describe one of the meanings of the expression 'to deceive oneself.'

2.2.3 'Self-Deception' and Biased Believing

Consider the following examples: 'He thinks I am going to visit her because he asked me to -- that is a typical case of self-deception;' 'Some students think that one can study the natural sciences without studying philosophy, but they are self-deceived;' 'Jones thinks that he will reach the shore -- how can one be so self-deceived.' I have a certain feeling that the meaning of these sentences has changed as compared to the instances when the expression 'to deceive oneself' is used to report that somebody's belief is unwarranted.

In order to show that there is difference between the use of 'self-deception' and that of 'deceiving oneself,' let me use another example. I am returning to the story about Brown,

White and Brown's wife who is unfaithful to Brown. I can imagine that Brown watches his wife leaving for the customary business trip and says to his friend White: 'You know, I still believe she is faithful to me, but probably I am just deceiving myself.' And the friend agrees: 'Probably you are.' Would one say in this situation that Brown is a self-deceiver or that he is in the state of self-deception? I think one would not. If one knows that Brown realizes that his belief may be incorrect, one would not say that Brown is a self-deceiver.[3] So White can say about Brown that Brown is deceiving himself, but he cannot say that Brown is self-deceived.

If I am right, the expression 'to deceive oneself' when it is used to designate belief in belief-adverse circumstances cannot be substituted with the term 'self-deception.' Canfield and Gustavson, Siegler and Scott-Taggart have described one of the meanings of the expression 'to deceive oneself' and not the meaning of the term 'self-deception.' Meanwhile, I do not want to say that the expression 'to deceive oneself' can never be used to replace the term 'self-deception.' If the mother of the boy who is cruel to animals makes herself believe what she knows is false, namely, that he is not cruel to animals, I could say about her both that she presents an instance of self-deception and that she is deceiving herself. Only in this case my claim about the person would imply more than just belief in belief-adverse circumstances. I am claiming that the mother has done something in order to have such unwarranted belief. Of course, it is hard to see the difference if one is presented with one sentence like 'Jones is deceiving himself.' It seems to me that one can tell which meaning is used in a particular sentence only if one knows or presupposes the context in which the phrase is ascribed to somebody.

---

[3] In the example, I used expression 'I still believe' to emphasize the possibility that one can believe something even if one entertains a thought that the belief could be false. The awareness of the possibility to be mistaken need not undermine the belief.

I think I have shown that 'to deceive oneself' can be used in a sense that is different from any meaning of the term 'self-deception.' Meanwhile, I have not shown yet what the difference is. Brown believes that his wife is still faithful while it is clear for everybody around that she is not. It certainly seems that unless Brown announces that he thinks his belief could be false, one could say that Brown is self-deceived, or that Brown is in the state of self-deception. The question is what one would mean by saying that Brown is self-deceived.

First, it seems that as in the cases when one uses the expression 'to deceive oneself,' one would ascribe to Brown a certain false belief in spite of the evidence of which Brown is aware, but does not recognize as evidence against his belief. It is more difficult to say whether one would claim that Brown is able of adjusting his belief to evidence. One would not say that Brown is stupid or mentally ill, so in some sense Brown is capable to adjust his beliefs. Nevertheless, it seems to me that when one says that Brown is self-deceived, one is claiming that something has gone seriously wrong with Brown's capacities to recognize the discrepancy between his belief and the evidence.

I also think that by ascribing to Brown self-deception, one is suggesting that there is some mental cause for Brown's incapacity to recognize the implications of evidence for his belief. When one is saying that Brown is self-deceived, one most likely thinks that Brown's wish that his wife would be faithful somehow influences his capacity to evaluate correctly the evidence. While in this case the influence is exerted by the wish, on other occasions, the belief may be influenced by something else. For example, Walter Raleigh writes about the men of Shakespeare's plays that "their imagination often masters and disables them" (175). He adds, "Self-deception, it would seem, is a male weakness" (175). Here some preconceived and imagined understanding of how things are undermines one's capacity to judge objectively. According to Raleigh, Macbeth "sees the murder as a single incident in the moving history of human woe" and fails to understand the practical aspects and consequences of his actions.

If Plato's use of the term 'self-deception' is to be related to some meaning the term has in the contemporary English, I think his interpretation must be mentioned here. Plato uses the term 'self-deception' to denote a certain failure of reasoning, and the preconceived and imagined understanding of how things are certainly can undermine one's capacity to judge things objectively and can result in a failure of reasoning. One can be carried away by one's thoughts and, interpreting evidence as supporting one's preconceived ideas, fail to notice the obvious.

I think I can try to define the meaning of 'self-deception' which the term has in ordinary language and which is not paradoxical. Sometimes when one ascribes to somebody self-deception, one is claiming that (1) the person is in a certain state of mind that causes the person falsely believe something that is not warranted by evidence, (2) the state of mind can be characterised as being biased by some wish, presupposition or interest, (3) and the person is not aware of the bias, and whenever one realizes that the belief is biased, one ceases to be in the state of self-deception.

One can notice the difference between this non-paradoxical understanding of 'self-deception' and the paradoxical, that is, making oneself believe something one knows is false. Neither the condition of 'making,' i.e., some action, nor the condition of two contradictory beliefs can be ascribed to the non-paradoxical meaning of 'self-deception' that shortly can be characterized as biased believing. The term 'self-deception' that is used to designate biased believing is also distinguishable from the expression 'to deceive oneself' that is used to designate unwarranted belief. The former designates a specific state of mind that is characterized by a wish or interest that causes the person to have a false belief, while the latter indicates only a discrepancy between one's belief and evidence one has. The meaning of 'deceive oneself' does not require the absence of awareness about one's state of mind, and it need not require the presence of interest or wish that biases the evidence. If students who falsely believe that there is no philosophy in the natural sciences are in the state of self-deception, they must want philosophy to be absent and their want should bias

the evidence. The students, of course, could be in such a state; nevertheless, I doubt that Derrida would claim that they are. More likely he is just saying that students err.

2.3 Other Alternatives to the Paradoxical Meaning

I have pinned down two meanings that 'self-deception' has in ordinary language. To describe them briefly, 'self-deception' can either mean a biased believing (the non-paradoxical meaning of 'self-deception') or making oneself believe what one knows is false (the paradoxical meaning of 'self-deception). I have also described some confusion with the definition of 'self-deception' as unwarranted belief. Now one could ask whether I have described all the meanings that 'self-deception' has. It, certainly, does not seem so.

For example, I can say that Fingarette does not accept the definition of 'self-deception' as making oneself to believe what one knows is false, because he claims that paradoxes arise from the characterization of self-deception in terms of belief and knowledge (*Self-Deception* 34). It seems that he does not accept the idea that 'self-deception' means nothing more than biased believing, because he claims that a self-deceiver persuades himself to believe contrary to the evidence and that "the self-deceiver purposefully brings it about that he is deceived" (*Self-Deception* 28,31). At the same time he says that "the self-deceiver is one who is in some way engaged in the world but who disavows the engagement, who will not acknowledge it even to himself as his" ("Self-Deception" 81). Should one understand this phrase as a definition of the concept 'self-deception'? If so, it is definitely not a definition that describes the usage of the word in ordinary language. I would be very surprised if somebody who is not familiar with philosophical analysis of self-deception would say: "What is self-deception? I don't know. Well, I guess it's a disavowal of one's engagement in the world." But if one allows Fingarette's definition to be a stipulative definition of the concept, then one can feel

embarrassed when asked about other definitions that do not match either of the two ordinary language definitions, neither the paradoxical nor the non-paradoxical one.

Is Audi's definition another stipulative definition? He says that a self-deceiver unconsciously knows some proposition, while sincerely avowing the negation of this proposition and the person has at least one want that explains why the person is in such a state (173). In ordinary language, people do not call anyone a self-deceiver meaning that the person unconsciously knows one thing but avows another. For example, one can compare the following versions of a statement about self-deception: 'Jones still believes that Mary will marry him, but he knows that she will not. How can one be so self-deceived?' and 'Jones still avows that Mary will marry him, but he unconsciously knows that she will not. How can one be so self-deceived?' I doubt that anybody asked to explain what he or she means by 'self-deception' would use the second interpretation. Nevertheless, it is very possible that if asked to explain how it is possible that one believes in one thing while knowing that the opposite is the case, one would claim that Jones unconsciously knows that Mary will not marry him. One could say that the ordinary language speaker assumed the unconscious knowledge in the first case but expressed the meaning imprecisely. I cannot provide conclusive evidence that such interpretation is wrong, but personally I believe that reference to unconscious knowledge is used as a way to explain the apparent paradox which a person recognizes only after the question about knowing not-*p* and believing *p* is asked. If one would ask the person what he or she means by unconscious knowledge, the most probable answer will be that there are better things to do in the world than answer silly questions. And I think that the inability to explain what one means by the term 'self-deception' does not imply that one does not know what 'self-deception' means, it rather implies that one does not really mean anything other than believing one thing and knowing the opposite at the same time. Only when challenged, does one realize the paradoxical nature of the interpretation.

So I would say that neither Fingarette nor Audi has presented the definition of the

meaning that the term 'self-deception' has in ordinary language. One can continue this list of definitions that will not match the definitions of ordinary language.[4] There are many who do not try to define self-deception and whose interpretation of 'self-deception' seems to be different from the two I have described. The reader should not misunderstand me. I am not claiming that the right way to define a concept is to provide the definition of the meaning that the concept has in ordinary language. Also, I am not trying to reject the definitions of philosophers that are not consistent with the meaning the term in ordinary language. I am simply trying to establish whether the definition some philosopher has provided corresponds to the meaning that the word has in ordinary language. If it does not, then the only conclusion is that the particular philosophers has his own understanding of what self-deception is. And my conclusion is that there are quite a few different definitions provided by philosophers none of which correspond to ordinary language.

I could conclude that some definitions of the concept 'self-deception' report the usage of the term 'self-deception' in ordinary language, and all the other definitions are different stipulative definitions. Unfortunately, such a position would leave me with some unanswered questions. If one will allow for many definitions to be stipulative definitions, then the discussion of self-deception seems to be impossible: everybody discusses something else. Meanwhile, it does not seem like the discussion of self-deception is completely incoherent. So what is going on? I think that in order to answer this question I must look again at Audi's analysis of self-deception.

Robert Audi starts his article "Self-Deception and Rationality" by listing a number of problems concerning self-deception that have puzzled philosophers. He announces, "This paper is based on the view that despite these difficulties the concept of self-deception is both explicable without paradox and useful in understanding persons" (169). He

---

[4] See, for example, Rorty's definition ("Deceptive Self" 25), or McLaughlin's (51-52), or Harold A. Sackeim and Ruben C. Gur's (150), or W. J. Talbott (30).

continues with a story about Othello and Iago as an example of interpersonal self-deception (170). As everybody knows, Iago deceived Othello. Now Audi wants to entertain a possibility that Othello is deceiving himself. According to Audi, Othello is attracted to Emilia, but, being a faithful husband of Desdemona, makes himself believe that he is not attracted and the attraction is gone. Audi announces that the example of Othello cannot be an example of self-deception because self-deception "apparently exhibits . . . both deceiver and deceived" and must include "a kind of duality" (171). Othello has none of these, his attraction and his belief that he is attracted to Emilia are gone. Audi continues by declaring that any account on self-deception must "speak to" the interpretation of self-deception that assumes that a self-deceiver believes something that he or she knows is not true, and Audi proposes his own account (172-173). Audi thinks that a person "is in self-deception" when he or she unconsciously knows some proposition, while sincerely avows the negation of this proposition and has at least one want that explains why the person is in such a state (173). Audi attempts to demonstrate the correctness of his account with a rather lengthy explanation of how Othello could be self-deceived (173-177). Audi writes,

> He [Othello] not only exhibits embarrassment around Emilia, but lavishes unusual attention on Desdemona at the earliest opportunity thereafter and protests too much both regarding his attraction to Desdemona and concerning his immunity to the charms of Emilia (175).

Audi remarks that his account "is meant to apply to paradigm cases, and it may not capture all the current admissible uses of 'self-deception'" (173). I am not sure what Audi means by 'admissible uses,' but his definition certainly does not capture any of the current uses of 'self-deception.' I have not read any other philosopher that would use an identical interpretation of 'self-deception,' and so one cannot say that his definition depicts 'self-deception' as it is used in philosophical discourse. And as I demonstrated, his definition does not reflect 'self-deception' in ordinary language either.

Audi's claims that self-deception 'apparently exhibits both deceiver and deceived' and presupposes 'a kind of duality' seem to suggest that he could subscribe to the usual paradoxical definition of 'self-deception,' namely, making oneself to believe what one knows is false. If that is true then it is not clear what Audi's definition defines. I doubt that he would say that he uses both 'self-deception' of ordinary language and a stipulative definition at the same time. It is clear that there is some connection between the two, but I doubt that Audi would say that by the word 'self-deception' he understands a state in which somebody knows *p* and believes not-*p* as well as the state in which one unconsciously knows *p* and consciously avows not-*p*. Rather he would say that if the term 'self-deception' makes some sense, it must be understood in terms of unconscious knowledge and sincere avowals.

So what happens to the ordinary language understanding of 'self-deception' of which Audi seems to be aware? Audi promises to analyse the concept and claims that he has done so. Nevertheless, it seems to me that meanwhile he has switched from an analysis of the concept to an explanation of self-deception, i.e., the phenomenon that is called 'self-deception,' and he claims that self-deception as a phenomenon is after all not paradoxical (172). His claim seems to be that the phenomenon of self-deception seems paradoxical at the first sight, but in fact is nothing more than unconscious knowledge and sincere avowals. Similarly one could switch from talking about the meaning of the words 'the sun rises' to talking about what really happens in the morning when 'the sun rises.' All that happens is that the earth rotates and the place where I stand is exposed to the sun.

How could such a switch happen unnoticed? I think that the reason for this is the fact that someone who knows how to use a word need not know how to state the meaning of it. Let me look at some examples. Amelie Oksenberg Rorty starts her article "The Deceptive Self: Liars, Layers, and Lairs" with a rather lengthy example of a cancer specialist who seem not to notice her symptoms of cancer and displays strange behaviours that suggest that she knows that she has cancer, for example, she writes a will. The

example is introduced with the words: "If anyone is ever self-deceived, Dr. Laetitia Androvna is that person" (12). A philosopher who denies the possibility of literal self-deception, Mary Haight, also knows when the word 'self-deception' must be used. She mentions, for example, a man who may have cancer, ignoring his symptoms or explaining them away, and a doting mother, blind to her son's faults in ways that do not really seem possible (vii). Audi also knows when the word is applied. When after providing the definition of 'self-deception,' Audi presents the example of Othello, he wants to provide an example that would show that his interpretation of 'self-deception' is possible. The interesting thing is that his Othello would indeed be called a self-deceiver in everyday language, even by those that have not read Audi's article and would not use the word in the sense that Audi does. So it seems that when philosophers explain what self-deception is they do not talk so much about the concept of self-deception as about something that they recognize being self-deception. I think that they recognize certain behaviour. Dr. Laetitia Androvna denies that she has cancer and seems to do this sincerely while some of her behaviour suggests that she knows she has cancer, for example, she writes her will. Othello avows that he is not interested in Emilia, but his behaviour suggests that he is interested in her and knows that. He 'exhibits embarrassment.' Should one say then that the term 'self-deception' is meant to designate certain behaviour?

I doubt that the word 'self-deception' means only certain kinds of behaviour. I would rather agree with Audi that the word 'self-deception' is an explanatory concept that implies an explanation of how certain behaviour is possible (189-19). When one claims that Jones is self-deceived, one does not just claim that Jones acts like he knows what he claims not to know, instead one claims that Jones in fact knows the truth. Of course, when Audi talks about such a concept he uses his own version of what 'self-deception' means, but it seems to me that 'self-deception' in ordinary language represents an explanatory concept. When I see that a mother claims that her son is a good boy and seems to do it sincerely, but I have good reasons to believe that she knows that her son is not a good boy, since she has

seen him killing bumblebees, I am explaining the behaviour of this mother. I am stating, correctly or wrongly, that she knows that the boy is bad, but makes herself to believe that he is not. And whenever it seems to me that I can explain in this way some behaviour (like ignoring or denying obvious things, or behaving in strange ways), I am saying that one has made oneself believe what one knows is false. I am saying that the person is self-deceived.

The good thing about this explanation is that it looks plausible. It looks plausible that the mother knows that her son is a bad boy and believes that he is good at the same time. The bad thing about this explanation is that it does not survive any analysis. While the words 'know,' 'believe' and 'make' stay as they are, the explanation seems meaningful. When philosophers try to analyse this explanation, they quickly get into trouble. For example, what do 'know' and 'believe' mean in this explanation? 'Consciously know' and 'consciously believe'? If the answer is 'yes,' how is such self-deception possible?

I should not be surprised about the obscurity of everyday language. I can remind the reader of the example that I used for showing the distinction between the meaning of a concept and the description of a phenomenon, namely, the phrase 'the sun rises.' Surprisingly, when I think about what exactly I mean by saying that the sun rises, I soon get into different kind of problems. The first interpretation that comes to my mind is 'the sun goes up.' Do I mean that the sun moves in the upward direction? Not really. I can try to define the meaning by saying that the distance between the sun and the horizon increases. Do I mean that the distance between the star Sun and the horizon increases? Not really. Do I want to describe my perceptual field and the position between the bright spot in my perceptual field, the sun, and the horizon? Maybe that is what I am doing, but the meaning of the phrase 'the sun rises' does not suggest this interpretation. All I am saying is that the sun 'goes up,' and I do not really think about what exactly this phrase means.

The problem with sentences like 'the sun is rising' and 'somebody believes what he or she knows is false' is that something convincing in these phrases. I can understand that somebody has reasons for saying that the sun rises. And I can understand that there are

some reasons why one wants to say that the person believes what he or she knows is false. Why does a cancer specialist deny that she has cancer when the evidence is obvious? Why does she write her will, if she thinks that she has no cancer? Why does Othello reacts so strangely when he is near Emilia? One way of dealing with the problem is to ascribe to the word 'know' a very broad or very specific meaning. As Paluch has indicated, philosophers like Freud and Demos do not talk about knowing in the ordinary sense of the word and talk about unconscious and latent knowing (270). The same is true of Audi, for example. Other philosophers would claim that my knowledge that *p* and my believing that not-*p* are somehow separated in my mind (see, for example, Rorty "Self-Deception" 130-131, King-Farlow 135, Davidson "Deception" 91-92, Sackeim and Gur 188, Pears "Goals" 76-77). Often philosophers would simply deny that the person really knows *p* when he or she believes not-*p* (see, for example, Paluch 275-276, Baghramian "Paradoxes" 172-173, Mele *Irrationality* 127, Siegler 471-472). And some would say that the self-deceiver does not really believe what he or she claims to believe (For example, Haight *A Study* 108, Kipp 261)

The question of whether self-deceivers know what they seem to know and whether their knowing is conscious or unconscious are only some of the questions that can be asked and are asked about self-deception. For example, one can ask whether the mother *makes* herself believe in her son's virtue. The ordinary language meaning of 'self-deception' seems to imply that the mother actively and intentionally chooses to believe one thing while knowing that the opposite is true. Of course, this description of what happens in cases of self-deception sounds paradoxical. How can one do anything like making oneself believe what one knows is false? Again, philosophers have taken different attitudes towards this question. Some philosophers deny any active biasing of beliefs (see, for example, Kipp 261, Johnson 152). And many suggest that the beliefs are biased but the biasing does not have the form of making oneself believe what one knows is false, unless there is some special sense of 'knowing' and 'believing' (For example, Davidson "Deception" 88,

Fingarette *Self-Deception* 47-48, McLaughlin 51, Mele *Irrationality* 127, Paluch 276).

It is also interesting to distinguish two different methods that philosophers use when they want to provide some explanation for behaviour that is associated with the term 'self-deception.' There are some philosophers who take concrete examples of self-deception and try to explain what happens in the mind of the person who ignores something obvious or displays a behaviour that suggests certain knowledge of the facts that the person denies (see Rorty "The Deceptive Self," Haight *A Study of Self-Deception,*). These philosophers usually start their analysis with particular examples of Johns and Marys who usually would be called self-deceivers, and the aim of the analysis is to show that the behaviour of these people, such as, denying obvious things, is explainable without using any paradoxical suggestions about knowing and not knowing or making one self-believe what cannot be believed. Other philosophers try to model what could happen in a self-deceiver's mind that would *resemble* these paradoxical interpretations (see, for example, Audi "Self-Deception and Rationality," Talbott "Intentional Self-Deception in a Single Coherent Self," Davidson "Deception and Division"). These philosophers usually concentrate on models of defective rationality.

Of course, the result of many interpretations and different approaches is different definitions of self-deception. But what is common among them? Why do all philosophers claim that they have defined self-deception? Why do so many different definitions come under one name? I think that the answer is simple. The common thing is philosophers' aim to explain the strange behaviour that usually associates with the ordinary language term 'self-deception.' They refuse the explanation of that behaviour suggested by the meaning of 'self-deception' as it is used in ordinary language, and try to substitute for it their own interpretation.

I think that I have been able to explain how the word 'self-deception' is used in ordinary language and why there are so many different definitions of the term in the philosophical literature on self-deception. I must say that the process of clarification has been

quite painstaking which implies the confusion in the use of the term. I think that this confusion is caused by various abuses of the ordinary language word in philosophical discourse. It seems to me that the analysis of different problems that come under the name of self-deception would be much clearer if philosophers would try to label these problems with their own names, such as, 'irrational behaviour' or 'motivated biasing of beliefs.' Personally, I will not follow my own advise and use the term, and the next chapter is meant to analyse the possibility of deliberate self-deception.

3.0 IS DELIBERATE SELF-DECEPTION POSSIBLE?

Before I proceed to answer this question, I think I need to clarify some of the concepts that I am using. In the first two sections of this chapter, I will try to explain my understanding of the concept 'deliberate self-deception' and the nature of the problem that my question is intended to present. I think that such an introduction will help the reader to understand better the analysis of the possibility of deliberate self-deception that I give in the remaining sections of the chapter.

In Section 3.1, I explain my understanding of deliberate action. Deliberate action is an action that is preceded by a state of intending. In Section 3.2, I present deliberate self-deception as a type of deliberate action. I confine the meanings of the term 'self-deception' to one meaning that the term has in ordinary language. I think that 'making oneself believe what one knows is false' is the meaning of 'self-deception' that could allow the possibility of self-deception as deliberate action and that is related to my interest in the possibility of controlling one's mind. I define deliberate self-deception as an action of making oneself believe what one knows is false that is preceded by intending to make oneself believe what one knows is false. The remaining sections are meant to analyse a possibility of such self-deception. In Section 3.3, I distinguish between two kinds of *deliberate* actions: basic actions and non-basic actions. A basic action is a deliberate action that does not need additional deliberate actions for its realization. Deliberate self-deception as a basic action consists of making oneself believe what one knows is false without intending any other action than making oneself believe what one knows is false, for example, without intending to forget evidence for something one knows. I argue that deliberate self-deception as a basic action is impossible, because there is no basic action of making oneself believe something, for example, believe that *p*. Since deliberate self-deception requires making oneself believe that *p*, there cannot be deliberate self-deception as a basic action. In Section 3.4, I consider the possibility of deliberate self-deception as a non-basic action, that is, as an action that requires for its realization some additional deliberate action. In the first part of the section (3.4.1), I show that in order to make oneself believe what one knows is false, one has to undermine one's explicit and conscious knowledge of the falsity of the proposition which one wants to believe. In the second part of the section (3.4.2), I examine the possibility of realizing deliberate self-deception by forgetting, which is a deliberate

non-basic action. Deliberate self-deception is making oneself believe what one knows is false; and in order to realize deliberate self-deception, the self-deceiver can try to undermine his or her knowledge of the falsity of some proposition *p* by trying to forget either that *p* is false or the evidence that supports the knowledge that *p* is false. In the third part of the section (3.4.3), I examine the possibility of realizing deliberate self-deception by deliberately reinterpreting the evidence for the falsity of *p*.

3.1 Deliberate and Intentional Actions

Usually, the adjective 'deliberate' specifies the nature of some action. For example, 'John deliberately stepped on the banana skin.' Sometimes the adjective 'deliberate' looks like it characterizes not an action but some state or event. For example, 'John made a deliberate error.' I think that such an expression is meant to describe the way the particular state or event has come about, namely, the action that has brought it about, and 'deliberate' here is not used to denote some intrinsic property of the state or event. At least, I cannot imagine what kind of intrinsic property that would be.

Usually, the adjective 'deliberate' can be replaced with the word 'intentional.' For example, 'John intentionally stepped on the banana skin.' For reasons that are unknown to me, philosophers prefer the word 'intentional,' which they sometimes substitute with the word 'deliberate.' I have not found a philosopher who would try to distinguish between the use of the terms 'intentional action' and 'deliberate action,' and in practice the use of these terms does not differ in any noticeable manner.[5]

The word 'deliberate,' according to the OED, originates from the Latin word *libra*, to balance. Here one could search for some differences between 'deliberate' and

[5] See, for example, Donald Davidson ("Deception" 86), Jerome A. Shaffer (78), Samuel Guttenplan (559), Lawrence H. Davis ("Action" 112-113).

'intentional' because the latter does not suggest anything that in any sense resembles balancing. Meanwhile, the verb 'to deliberate' usually is explained as a weighing of reasons or evidence. So it could be tempting to suggest that 'deliberate action' is an action that follows, or results from, deliberation. Since not all intentional actions are preceded by deliberation, one could use the condition of deliberation to distinguish between 'deliberate' and 'intentional' actions. For example, I suddenly notice a silverfish running on my bathroom floor and I step on it with the purpose of smashing it -- I have stepped on it intentionally; but since before my action I did not weigh the reasons for and against my stepping on the silverfish, I have not performed a deliberate action. Of course, the distinction I just portrayed is concocted and does not correspond to the way the terms 'intentional action' and 'deliberate action' are used in philosophical literature or ordinary language. Both in philosophical literature and ordinary language the action I describe would be called a deliberate action. I saw the silverfish and intentionally stepped on it; and even if I did not deliberate on my future action, I did it deliberately.

The only difference I can notice between the uses of 'intentional' and 'deliberate' is that in ordinary language the word 'deliberate' is usually ascribed to intentional actions that are considered to be condemnable, and usually such actions are condemnable because they are intentional. Thus, to accuse me of an act of cruelty, someone would, first of all, insist that I stepped on the silverfish deliberately; while to defend myself against such accusations, I would claim that I stepped on it unintentionally or by accident.

I doubt that philosophers think about blameworthy actions when they write about deliberate control or deliberate coughing (Shaffer 78, Guttenplan 559). In the philosophy of mind, the aspect of blame is somehow lost. My suggestion for an analysis of possibility of deliberate self-deception is not meant to imply blameworthy actions either. So I propose to ignore this aspect of the meaning. But before I explain what I call a deliberate action and deliberate self-deception, I want to say some words about the ways philosophers have interpreted the notion of intentional action. Thus, I will show the reasons why I want to

distinguish between intentional and deliberate action.

It is quite easy to separate those events that come under the name 'intentional actions' and those that do not. The sun shines -- that is not an intentional action, or action at all. Mary runs after a ladybird -- that is an intentional action. John does not know that there is a banana skin on the floor and slips on this skin -- that is not an intentional action. As my analysis of 'self-deception' has demonstrated, knowledge of how the word is used does not guarantee that it will be easy to formulate the meaning of the word, in this case, to state what intentional action is and what makes it different from other events in the world.

Quite a few philosophers have tried to elucidate the notion of intentional action, and I will not attempt here to give an account of everything that is said about actions, intentions and intentional actions. I want to mention only one aspect of the discussion on the nature of intentional actions, namely, the difference between intentional actions and actions that are intended. As G. E. M. Anscombe in her book *Intention* notes, there is a certain temptation to say that the words 'intention' and 'intentional' mean different things in different contexts (1). When one talks about 'intentional actions' or 'intentions in actions,' one thing is meant; when one talks about intentions as certain states of mind, the word 'intention' is used somehow differently; and when one talks about intentions that concern the future actions, the word 'intention' has some specific meaning different from the other uses.

For example, Mary had planned to spend her holiday running after ladybirds and so she did. It seems natural to say that Mary had an intention in the form of a plan or idea that she later realized. Nevertheless, not all actions that are called 'intentional actions' are intended beforehand. First of all, there are intentional actions that are performed spontaneously (Searle 84). For example, if a plate slips from my hands and I catch it, my catching of the plate is an intentional action that is not intended beforehand. Or there are intentional actions that are performed out of habit. For example, Seth went to the medicine chest for some aspirin and, instead of aspirin, absent-mindedly took the tooth-paste out of

the chest (Davis *Theory* 59). In these cases, the action is intentional, because Seth did not just grab the first thing that happened to be in the chest; he forget that he intended to take some aspirin and absentmindedly took something that he was used to taking out of the medicine chest. There is no prior intention to take tooth-paste, and the word 'intention' must be used in some other sense.

Anscombe claims that the word 'intention' is used in the same sense both in the case of an action that is intended in advance and in the case of an action that is not intended in advance (90). Intended or not, intentional actions, according to Anscombe, are actions that can be explained by reasons on which the person acts (9,90). And the word 'intentional' refers to a certain form of description of actions, namely, the description that indicates the reasons for these actions (84-85).

Several philosophers have argued that the intentions that concern future actions are not just descriptions of the reasons for such actions, and they distinguish between intention, or intending, as a particular state of an agent, or mind, or consciousness, that sometimes precedes the action and intention as an intrinsic characteristic of intentional action (Davis *Theory* 59-60, Davidson "Intending" 84-85, Searle 84-85). Davis claims that there are states of intending and there are intentional actions, and the two must be "sharply distinguished" (59). Davidson talks about 'pure intending,' and, according to Davidson, pure intending is not always present when somebody acts intentionally ("Intending" 88). Searle distinguishes between 'intention in action' and 'prior intention,' and there are intentional actions that are not preceded by prior intention (84).

If there are such states as pure intending or prior intention, I would like to know more about their characteristics. Davis describes a prior intention in the following way: (1) "If *x* intends to do an A [an action], then *x* believes and would claim to know that he will intentionally do an A, or at least try" (*Theory* 76); (2) "If *x* nonobservationally believes and would claim to know that he will intentionally do an A, or at least try, then he intends to do an A (or at least to try)" (*Theory* 77); (3) "Intending to do an A is just nonobservationally

believing (and being ready to claim knowledge) that one will do an A, or at least try" (*Theory* 77). The problem with the attempts to identify intention with the beliefs about one's future action is that one cannot distinguish between wishful thinking and intention, or simple prediction of one's actions and intention (Bratman 377). For example, knowing how much coffee I usually drink, I can predict that at Mary's party I will drink four cups of coffee. Nevertheless, my prediction does not imply that I intend to drink four cups of coffee. I can intend to drink twelve cups at Mary's party, but after the fourth, I have a definite feeling that I do not want more coffee. In this case my prediction that I will drink four cups would be true, but my intention to drink twelve cups would not be carried out.

According to Searle, intending is one of the Intentional states (3).[6] Intentional states are mental states that are "directed at or about or of objects and states of affairs in the world" (1). To illustrate what Searle means, I can say that believing and being angry are mental states about something, for example, about bad weather, while pain and bad mood are mental states that are not about or directed at anything. I am angry *about* bad weather, but my bad moods is not *about* bad weather, it is rather *caused* by bad weather. Intentional states are different, and Searle uses several criteria for distinguishing them. Two of the most important ones for understanding any kind of intention are the psychological modes and the Intentional content. Intentional content describes what the Intentional state is about, or at what it is directed. So if I believe that it is raining, the content of my mental state is that it is raining. Searle does not say very much about the psychological mode of Intentional states, but basically it is the way the Intentional content is presented in the Intentional state. Thus, my believing that it is raining differs from my being angry that it rains, or my being glad that it is raining.

---

[6] In order to distinguish between the word 'intentional' that is used to describe a particular kind of action and the word 'intentional' that is used to characterize a property of many mental states, Searle capitalizes the latter.

According to Searle, there are two Intentional states that are called intentions. One is intention in action; the other one is prior intention (84). Searle is quite clear about the differences in the content of these Intentional states. The intention in action is directed at some *event*, for example, a movement of my arm, while the prior intention is directed at *action*, for example, the action of moving my arm (92-93). He also is quite clear about the psychological mode of intention in action. According to Searle, there is a certain experience of acting (87). If I have the experience and the arm goes up then my intention in action is carried out, or satisfied, and I have performed an action. If I have the experience of lifting my arm and the arm does not go up, my intention is not satisfied and there is no action. To illustrate his claim, Searle describes an experiment with a patient whose arm is anaesthetized and who is asked to raise this arm. "The patient's eyes are closed and unknown to him his arm is held to prevent it from moving. When he opens his eyes he is surprised to find that he has not raised his arm" (89). Similarly, if my arm goes up and I do not have the experience of acting, there is no intention and there is no action. Searle mentions an example of a patient whose arm moves because of the electrode applied to a particular part of his brain. The patient denies that he has moved the arm (89).

Unfortunately, it is not so easy to characterize prior intention. Searle shows the distinction between the Intentional content of both kinds of intentions, but he does not characterize the mode of the prior intention. A prior intention is directed at the action, and Searle explains that in order for a prior intention to be carried out, or satisfied, the person must act, for example, there must be a certain experience of lifting my arm and the event of my arm going up. What is not so clear is the character of the psychological mode which the prior intention has. Searle claims that both intention in action and prior intention are causally self-referential, that is to say, the carrying out of the intention requires not just that some movement or action follows the intention, but that the movement or action is caused by the intention. So maybe the difference in the mode is the difference between the experience of causing some event (causing the movement of my arm) and the experience of

causing an action (causing the movement of the arm together with the experience of moving it). Nevertheless, as Davidson has noted, there can be intentions that are not followed by action ("Intending" 84). Describing intentions in action, Searle himself presents a case characterized by intention in action but lacking an action: I experience my arm going up, but my arm does not go up (89). Here, the psychological mode of intention is the experience of moving one's arm. Searle has not described a similar example of prior intention, and it is not clear what it would be like. An experience of an action that is not followed by the action is hardly a good characterization of prior intention, because the prior intention is not an experience of action. My prior intention to collect butterflies is not an experience of collecting butterflies; it is something else, and I want to know what it is.

Searle also suggests that there are certain similarities between the two kinds of intentions and perception and memory. A memory of seeing a flower represents the experience of seeing and the flower, and a prior intention represents the experience of acting and certain movement (95). The emphasis lies on the phrase 'to represent.' According to Searle, the word 'to represent' is used to describe the conditions under which the particular Intentional state is 'satisfied'(12). For example, a belief that *p* is satisfied when *p* is true. Similarly, a state of memory is satisfied when there has been the visual experience of a flower that is caused by the presence of the flower (95). Prior intention seems to be satisfied when this intention is followed by an action, that is, experience of acting and the movement of the arm, for instance.

I must say that this analogy is not very helpful for understanding the psychological mode of intention. To know what memory is, it is not enough to know that memory concerns past experiences. My anger or joy can concern my past experiences, too, and the characteristic that separates these different Intentional attitudes is their psychological mode. Similarly, to know what prior intention is, it is not enough to know that prior intentions concern actions. I want to know what makes the intention to lift my arm a different Intentional state from, for example, imagining that I lift my arm. In conclusion, I can say

that Searle does not provide a satisfactory characterization of prior intending and he does not show the way one could attempt to characterize the psychological mode of this intention.

I think that of the three philosophers I have looked at, Donald Davidson characterizes intending best of all. In his article "Intending," Davidson claims that intending, or pure intending, is an all-out judgement. I must add immediately that this claim that intending is a judgement should not be taken at face value. Davidson claims that he analyses judgements in order to "mark differences among the attitudes" (97). Davidson gives this comment only in a footnote, it seems to me that this comment is important. The main distinction between judgement and attitude seems to be the propositional form that judgements necessarily have, but attitudes need not have. According to Davidson, there are judgements that correspond to certain attitudes. For example, a judgement about the desirability of something corresponds to the attitude of wanting (96). It seems that the all-out judgement corresponds to the attitude of intending. Nevertheless, Davidson in one paragraph claims that intention *is* a judgement (99), but in another he states that "intending and wanting belong to the same genus of pro attitudes *expressed* [my italics] by value judgements" (102). So it is not clear how seriously one must take his claim about intending being a judgement.

Davidson's interpretation of pure intending is easier to understand when one knows Davidson's objections to the view that pure intending is nothing but wanting to do the action in question. As Davidson correctly argues, the judgement that something is desirable does not mean that I intend to do the particular action. For example, if I conclude that despite the danger of being bitten by a tsetse fly it is desirable to visit Africa, that does not mean I intend to go to Africa. Davidson writes,

> It is a reason for acting that the action is believed to have some desirable characteristic, but the fact that the action is performed represents a further judgement that the desirable characteristic was enough to act on (98).

According to Davidson, the further judgement is 'all-out judgement' or pure intention (99). The term 'all-out judgement' sounds a bit unusual, and I think that a better expression to characterize this judgement and ultimately the state of the agent is 'commitment to action' which is used by Bratman (Bratman 376). As Davidson says, there can be different beliefs and desires that precede the intention and influence what kind of intention it will be, but at one moment there must be some commitment to one particular action. As an all-out judgement differs from other judgements, so commitment differs from other attitudes or states of the agent. It is also clear that one can have a commitment to some action without really experiencing or carrying out this action.

There is just one more comment that I want to make about pure intending. Davidson emphasizes that one can have pure intending to do some action "without having decided to do it, deliberated about it, formed an intention to do it, or reasoned about it" ("Intending" 84). I must agree that commitment to some action need not require previous deliberation on whether or not to commit oneself to this action or not. Intentions can sometimes be quite spontaneous. I can agree that intending may not be preceded by decision as long as the latter implies making a choice among options. I can also agree that intentions are not always expressed in clearly articulated form like, for example, "I intend to shut the window." Nevertheless, Davidson's claim may sound mysterious because Davidson's emphasis on the connection between intending and the judgement suggests an explicitly stated intention, so in some sense a 'formed' intention.

It seems that the problem lies in a specific use of concepts. First, as I already noted, Davidson chooses to analyse judgements only because he thinks that the analysis of them will help him to mark differences among different attitudes. If, after all, pure intending is an attitude, it does need to have the form of an explicitly stated judgement. Secondly, Davidson's claim that pure intending does not require a 'formed intention' is not meant to suggest that pure intention is something vague and amorphous or that it is somehow reducible to something else, for example, beliefs. According to Davidson, the intention is

not formed as far as "forming an intention requires conscious deliberation or decision" ("Intending" 89). It seems that the only thing Davidson wants to say is that to have a pure intention, one does not need to engage in a process of conscious weighing of reasons pro and contra an action.

I can summarize the results of the analysis of different accounts on intended actions. Not all, but some actions are preceded by an attitude that is called intending. The intending that precedes an action is a commitment to that action. Intending is a specific attitude different from the desire to act in a certain way or the belief that one will act in a certain way. Intending can be the result of deliberation or the comparison of different options, but neither deliberation nor weighing of options is necessary for intending.

Finally, I am ready to define what is deliberate action. Deliberate action is an action that is preceded by intention, or a commitment to this action. So deliberate self-deception is an action that is preceded by intention to deceive oneself. Not all intentional actions are intended beforehand. For example, spontaneous actions are not intended, but still count as intentional actions. So, an account that explained self-deception as a spontaneous avoidance of evidence should be considered as an account of intentional but not deliberate self-deception.

3.2 Deliberate Self-Deception

In Chapter 2 of my thesis, I tried to show that in ordinary language the word 'self-deception' is used in two different ways. 'Self-deception' is understood both as making oneself to believe what one knows is false and as biased believing. The former formulation is paradoxical, the latter is not. The second of the two is understood as a state of mind in which one finds oneself rather than brings it about. My initial interest in self-deception concerned the possibility of controlling one's mind and making myself believe what I am disinclined to believe, and the interpretation of self-deception as making oneself believe

what one knows is false seems to fit better to my initial interest than self-deception interpreted as a state that is not brought about by the action of the agent. So I will concentrate on the interpretation that suggests certain actions on the part of the agent, that is, self-deception as making oneself believe what one knows is false.

The notion of making oneself believe what one knows is false has puzzled philosophers because of its paradoxical nature. Dealing with the paradoxical aspects of the meaning of the term, philosophers have adopted two strategies. One strategy is to identify situations in which the paradoxical term 'self-deception' is used and to explain the behaviour of so-called self-deceivers in a way that does not contain paradoxical accounts of the self-deceiver's states of mind or intentions. Philosophers who choose this strategy usually try to provide their own, non-paradoxical, definition of self-deception. The second strategy in dealing with paradoxical 'self-deception' is to provide some model of what could happen in the human mind that would *to a certain degree* correspond to the paradoxical formulation of self-deception.

No doubt, the explanation of the behaviour of the so-called self-deceiver can demonstrate the reasons why one is tempted to use a paradoxical called the cases of self-deception, and modelling can, and is meant to, provide an adequate explanation of real examples, so the two methods are compatible and do not exclude each other. Nevertheless, at least sometimes examples of self-deception can be explained in a simpler and more plausible way than by models of paradoxical thinking processes or paradoxical states of mind. For example, Mary ignores an obvious fact that her husband is unfaithful. If there are good reasons for claiming that Mary simply is too busy to notice the evidence of her husband's unfaithfulness and she does not really make herself believe what she knows is false, then there is no need to evoke any models of paradoxical thinking. So models of possible self-deception should not be perceived as the right way of explaining the behaviour that is usually associated with the name 'self-deception.' Even if I prove that Mary could have made herself believe what she knew is false, I cannot claim that Mary has

indeed made herself believe what she knew is false, unless I have shown that other explanations of Mary's ignoring of the obvious facts are wrong.

My interest in self-deception arises from my interest in the possibility to control what one believes and what one does not believe. If self-deception sometimes is the deliberate making of oneself to believe what one knows is false, I would have found one of the ways in which one can control what one believes. While philosophers sometimes charge self-deceivers with intentional self-deception, it is hard to tell to what extent self-deception could be intentional.[7] No doubt, the fully conscious decision to make oneself believe what one knows is false is an extreme and the most implausible version of intentional self-deception. Implausible does not mean impossible, and I could claim that some self-deceivers are consciously controlling their beliefs. The interesting thing about this claim is that it would be very hard to prove or disprove such a claim by observing the behaviour of some self-deceiver. If Mary believes in something that is obviously false, how could I tell by observing her behaviour whether she is deliberately deceiving herself, deceiving herself intentionally but not deliberately, or her mind has played some nasty trick on her? I doubt that it is possible to tell which of the three is the case. Of course, I could adopt the explanation that seems to me more plausible, but this approach would still leave a possibility that another explanation is the correct one. For example, I can say that Mary simply pretends to ignore the fact that her husband is unfaithful to her, but unless I know that deliberate self-deception is impossible, there is a chance that, knowing that her husband is unfaithful to her, Mary makes herself believe that he is faithful and makes it deliberately.

To find out whether deliberate self-deception is possible, I want to choose the second of the methods I mentioned in the beginning of this section, namely, I will try to

---

[7] Some philosophers who have suggested intentional self-deception: Mele (*Irrationality* 133), John King-Farlow (132-133), Jennifer Radden (115), Mary Baghramian ("Strategies" 93).

find out whether there is a model for deliberate self-deception. I will try to analyze the possibility that self-deceivers intend to make themselves believe what they know is false and ultimately succeed in their attempt. I do not claim that my analyses will explain all cases of what is called 'self-deception,' but I think that this analysis will shed some light on what can and what cannot be claimed concerning self-deceivers. I think that this analysis particularly concerns the claims that self-deceivers choose to be deceived, or choose to believe what they know to be false, or choose to believe what they want to believe. I want to know whether it is possible to deceive oneself deliberately.

3.3 Self-Deception as Basic Action

Some of my intentions are easier to realize than others. If I intend to touch my ear, I can realize my intention without delay; if I intend to catch a dragon-fly, there are many things I must do before I catch one, for example, I have to leave my office because there are no dragon-flies in it; if I intend to think of a sentence with the subject 'dragon-fly,' I can produce one 'just like that': 'A dragon-fly flies;' if I intend to write a poem such that the end of each line rhymes with the word 'ear,' I would have to think for a while before I could come up with one.

I think that the difference between such 'simple' and 'complicated' actions is well formulated by John Searle. He distinguishes between basic actions and actions that are not basic. Searle defines basic actions, or more precisely -- a basic action type, the following way: "*A* is a basic action type for an agent *S* iff *S* is able to perform acts of type *A* and *S* can intend to do an act of type *A* without intending to do any other action by means of which he intends to do *A*" (100). If I understand Searle's formulation correctly, the word 'intend' is meant to designate a prior intention, and the basic action is a deliberate action. So, for example, turning on the light is a basic action, because I do not intend to turn on the light *and* to reach for the switch *and* to turn the switch: I intend to turn on the light and just reach

for the switch and turn it. Of course, there can be circumstances when I cannot realize my intention of turning on the light without intending to do something more than just turning on the light. For example, it could be dark in the room and I might not know exactly where the switch is located. Therefore, to carry out my intention to turn on the light, I first of all intend to find the switch. In this case, turning on the light is not a basic action. As Searle says, his definition of basic action type makes actions basic relative to the agent and his or her skills (100).

The name 'basic action' is not invented by Searle, but his understanding of basic actions is different from, for example, that of Arthur Danto, who introduced the term into the philosophical discourse, or Alvin I. Goldman, who discusses basic act-types and act-tokens. Danto defined basic action as an action that is not caused by any other action (Danto 142). Goldman's definition of basic action-type requires for the action to be the result of a want, and it requires that basic action-types do not depend on the knowledge of how the act must be performed and knowledge about causal laws that would produce the desired action (Goldman 66-67). The most important difference between these two and Searle's definition is that neither Goldman's nor Danto's definition of basic actions includes the condition of intending.

I think Searle's distinction between basic actions and actions that are not basic reveals a real distinction between two different types of intended actions. The distinction is important for me because I am interested in actions that are intended and carried out according to this intention, that is, I am interested in deliberate actions. Moreover, his definition is helpful for formulating one type of action that is noticed by philosophers writing about the possibility for one to believe at will. For example, Bernard Williams writes that some things happen to respond to the will and some not (148). To make oneself blush, one can use roundabout routes like placing oneself in a situation which would make one blush, but one cannot blush at will (148). Similarly, one could make oneself believe in something by going to a hypnotist and acquire a belief by suggestion, but one cannot

believe at will (149). Jon Elster claims that some people try to achieve "by one stroke and at will what can at most be realized at one or more removes" (*Sour* 56). Though one can try to fall asleep by trying to distract oneself from any thoughts, one cannot make oneself fall asleep at will or make oneself believe what one wants to believe (*Sour* 45,52). I think that Searle's definition of basic action can provide a good formulation for what Williams and Elster want to say. Both Williams and Elster claim that believing as a basic action is impossible, that is to say, I cannot intend to make myself believe that *p* and carry out my intention 'just like that' without intending to do anything more.

If a basic action of making oneself believe that *p* is impossible, then the same is certainly true about the basic action of making oneself believe what one knows is false. In order to achieve the latter, one must be able to realize the former. For a while, I will leave aside the aspect of one's knowing that *p* is false and will address only the question of the possibility of making oneself believe that *p*. The claim that making oneself believe that *p* cannot be a basic action seems to be correct, but I would like to explain why it is correct. To avoid repeating the formula 'a basic action of making oneself believe that *p*,' I will use Williams' and Elster's expression and call this kind of basic action believing at will, or believing that *p* at will. One of the comments one could make about the claim that believing at will is impossible is that nobody really claims that such an action is possible. In other words, there is nothing interesting in the claims about the impossibility of such an action. I must agree that I have not heard anybody saying that he or she can believe at will whatever and whenever he or she wants. At the same time, there are indications that believing at will is considered a possible action. For example, H. H. Price turns his attention to some expressions in ordinary language that seem to imply there is believing at will. For example, he employs the phrases "I prefer to believe that," or "I can't and won't believe this," or "I refuse to believe that" (3,7,11). It seems that the words 'prefer,' 'can' and 'will' imply there is a choice on the part of the person. The person can choose to believe one or another statement.

Remember also Elster's claim that "it can hardly be denied that people often try to achieve by one stroke and at will what can at most be realized at one or more removes" (*Sour* 56). He does not explain which things one can realize 'at several removes' but cannot realize at will. Nevertheless, among his usual examples of attempts to bring certain states of mind 'at several moves' is the attempt to make oneself believe that *p* (*Sour* 57). So most likely Elster would allow that somebody could attempt believing at will. When Elster characterizes different kinds of self-deception, he says, "At one extreme is the attempt to carry out, say, the decision to believe in a direct and fully conscious manner" (*Ulysses* 176). Elster does not say that attempting to believe at will implies that the person considers believing at will possible, but I would certainly say that if there are such attempts, then those making them must also be the persons who claim that believing at will is possible. It is not an evident truth that there is a connection between attempting to believe at will and believing that believing at will is possible. Therefore, I will mention some considerations on this matter.

As Davidson has indicated, it is impossible for me to decide to do something, if I believe that the action I want to bring about is impossible (Davidson "Intending" 93). I think that this claim is true also for intending. For example, I cannot intend to flap my ears, because I know that it is impossible. I cannot even say that I intended but failed to carry out my intention, because intending involves certain commitment to the action that I decide to perform, and this commitment is absent when I know that the action is impossible. I can consider in my mind a thought 'I will flap my ears,' but this entertainment of a thought is not intending yet. So it seems that I cannot really intend to do something that I think is impossible.

Besides actions that are believed to be definitely possible and actions that are believed to be impossible, there are actions about which one does not know whether they are possible to realize or not. For example, I may not know whether it is possible to believe at will. So one could say that in these cases one can *try* to realize the action. It is certainly

true that I do not know whether it is possible to swim three minutes under water without breathing and still I could try to do that. Nevertheless, the question is whether, in this case, my trying implies any attitude towards the possibility of the action. I think that it does. At least, whichever basic or non-basic action I think about, I discover that if I intend to do something, I do not need to know whether the action is possible, but I certainly have to believe that it might be possible. I cannot intend to chirr like a grasshopper unless I believe that it is somehow possible to do that. I cannot intend to swim like a water-measurer unless I believe I might be able to do that.

It seems to me that if I can try to believe at will, I must believe that such an action might be possible. Nevertheless, when I think about the possibility of believing at will, the first problem that strikes me is that it is not clear what it is like to believe at will. Believing in itself is not an action. While the realization of my intention to make myself believe that *p* certainly would be an action, it is not clear what kind of action it is. In cases when I do not know whether the action is possible and believe that it might be, I usually know what I should do in order to realize it or try to realize it. In the case of believing at will, I am simply left with nothing that I could try to do. The same is true of Elster's example of making oneself fall asleep (45). If falling asleep must be understood as something that people try to achieve at one stroke and at will, then it is not clear what the action could be, what they could try to do. I can imagine different methods for making myself fall asleep, such as counting lambs and trying to think about something relaxing, but all these methods indicate that falling asleep is not a basic action and cannot be carried out just like that. Both in cases of believing and falling asleep, the basic action is impossible to realize just because there is no basic action that could be called 'believing' or 'falling asleep.'

There is one more place in which one can look for possibility to believe at will. It seems true that I am not aware of all the actions I can do. For example, I do not know whether I can jump over a fence two meters high. Most likely I cannot, but there could be a situation, fleeing from a bear for example, when I suddenly jump over this fence.

Something similar is suggested by William James in his article "The Will to Believe." He denies that one can believe whatever one wants. For example, one cannot imagine that two one-dollar bills in a pocket make hundred dollars (5). At the same time, James seems to suggest that in particular situations, one could make a choice between believing one thing or another (3,11). He characterizes this situation as a situation of genuine option which consists of three elements (3). First, the person who chooses between two hypotheses must consider both hypotheses as being 'alive,' that is, they must have some appeal to the person. Second, the hypotheses must be forced, that is to say, the person cannot avoid choosing between the two options. Third, the choice between two hypotheses must be "momentous" (3). The third clause means that the situation of choice is unique and has important implications for the person that makes the choice.

Unfortunately, the situation of a genuine option does not generate believing at will. First of all, it is hard to see how anybody ever could be forced to make a choice between believing two opposite beliefs. Of course, asked whether I believe that trees have leaves or that they do not I answer that I believe that they do. I doubt that I make any choice between believing that trees have leaves and that they do not, and I certainly do not believe at will that trees have leaves. I just believe that trees have leaves. One could object that the option between believing that trees have or have not leaves is not an option that is 'alive' and that in case of options that are 'alive' one has to make a choice. I must say that even in cases when both options have intellectual appeal, I am not forced to make the choice between believing one of the opposites. In a situation when both options seem to me equally believable, I simply do not have beliefs about the truth or falsity of the opposites. I am not making a choice, or believing at will. If it seems plausible both that my son is cruel to animals and that he is not, I can look for evidence that would support one of the possibilities, but I am not forced to make a choice between believing that he is cruel to animals and that he is not.

One could still insist that the options I just presented are not genuine options, and

when one is confronted with genuine options, one makes an act of choice that could be called believing at will. My only concern about this claim is that I cannot imagine any such options. At least, James has not provided an example of a situation when one is forced to choose between believing in one of two opposite propositions, or an example of somebody who makes this choice.

For example, he claims that the "question of having moral beliefs at all or not having them is decided by our will" or that questions of personal relationships cannot wait for the answer because waiting can cause the failure in relationships (22-23). Finally, he claims that the questions of faith must be decided by one's "active good-will" (28). I think that in all the three cases the choice compels one to act in one or another way, but never to believe one of the proposed beliefs. I can agree with James that moral questions are questions "whose solution cannot wait for sensible proof" (22). Nevertheless, I doubt that moral questions require the choice between beliefs. A doctor can be forced to choose between doing an operation of abortion or not doing, but he or she cannot be forced to choose between believing that abortion is morally permissible or believing that it is not. While the doctor has no options except to do the operation or not to do, there is always a possibility to do or not to do the operation without believing that abortion is morally right or wrong. The doctor can suddenly feel a conviction that abortion is wrong, and it is possible that after reflecting upon the arguments supporting or opposing abortion, he or she comes to the conclusion that abortion is morally right -- in neither case would I be willing to say that the doctor made his or her choice at will, and I cannot explain what it would be like to make such a choice at will. I would say the same about one's choice of faith. One can be forced to choose between living according to the Commandments and not living according to them, between going to church and not going, but one cannot be forced to believe in any religious doctrine or choose to believe in one at will.

James also describes a situation when a person is forced to answer the question 'Do you like me or not?' I can agree with James that in certain situations the hesitation to

answer this question and looking for the correct answer can have unpleasant consequences. Still, I would not consider this example as an example of forced choice between believing that one likes some person or that one does not. The only choice that is forced here is the choice between actions: either to say 'Yes,' or to say 'No,' or to be thoughtfully silent. Of course, it is possible that one answers sincerely after thinking about the question or right on the spot, but it is hard to see how the answer could be a result of willing to believe, or intending to believe, or deciding to believe.

It is very possible that James is claiming just that sometimes one has to make decisions on the basis of feelings and not on the basis of justified belief. One cannot think very long after the question about one's liking or not liking and one has to answer immediately relying on one's feelings. If James does not want to say anything more than that, then his use of the word 'will' is somewhat puzzling, and he certainly does not explain how believing at will is possible. I think I must conclude that none of the philosophers have told me how deliberately making oneself to believe some *p* could be possible. Consequently, it seems to me that deliberate self-deception as a basic action is impossible. If there is deliberate self-deception, it cannot be realized with just one intention, or 'just like that.'

3.4 Self-Deception as Non-Basic Action

Since deliberate self-deception as basic action seems to be impossible, the only place to look for deliberate self-deception is among actions that are not basic, namely, the actions that require for their realization some additional intended actions.

3.4.1 The Condition of Knowing

It is certainly hard to tell which actions are those that one needs to carry out in order

to realize one's intention to make oneself believe what one knows is false, that is to say, to realize deliberate self-deception. To specify the boundaries in which the needed actions could be found, I want to look at the implications of the condition that one should know that the belief one makes oneself believe is false.

I must say that the implications are quite discouraging. The problem lies in the nature of beliefs. It seems that when I claim that I believe that *p*, I am claiming that *p* is in fact true. Using Williams's expression, I can say that "beliefs aim at truth" (136). Consequently, if one at the same time claims both that one knows that *p* is false and that one believes that *p*, one asserts a contradiction. Such a claim would mean that one at the same time consciously and explicitly believes something that one consciously and explicitly believes to be false. Since beliefs aim at truth, such a state of believing is paradoxical.

I think that I can also claim that such a state of believing is impossible. At least, any attempt to imagine a state of consciously and explicitly believing in something that one consciously and explicitly believes to be false fails. I would discount any reports about believing something one knows is false as misuse of language, or as report that employs some specific understanding of the words 'believing' and 'knowing.' Davidson writes that "nothing a person could say or do would count as good enough grounds for the attribution of a straightforwardly and obviously contradictory belief" ("Deception" 81). I think that the same is true about the attribution of consciously and explicitly believing something one consciously and explicitly knows to be false. If one announced that one is believing something one knows is false, I would first of all try to understand what one means by the words 'knowing' and 'believing.' For example, if one said that one knows that one's friend is guilty of a crime, but one still believes that the friend is innocent, I would interpret this claim as suggesting doubts about the guilt or innocence of the friend, or failure to come to terms with the deeds of one's friend. Certainly, I would hesitate to ascribe to the person a paradoxical state of believing something the person knows is false.

If self-deception were a state when a person consciously and explicitly believes something one consciously and explicitly knows to be false, self-deception would be an impossible state to achieve. Similarly, a thought about intending to bring about such a state could be entertained, but nobody could really intend to achieve this state. Nevertheless, the question I intended to answer was not the question of whether one can bring about a state of believing consciously and explicitly what one knows is false, but the question of whether it is possible to make oneself believe what one knows is false and to do it deliberately. Intending to make oneself believe what one knows is false is not identical with intending to believe consciously and explicitly what one consciously and explicitly knows is false. For example, one can easily imagine reasons for the former, while it is hard to imagine why anyone should attempt the latter. I can try to make myself believe what I know is false, because I find the truth disturbing and I prefer illusions to the truth. Nevertheless, the only reason to think about intending to bring about the state of consciously and explicitly believing and knowing opposite things would be a whim or curiosity about one's mental capacities.

I think that, knowing the reasons why one would deliberate deceive oneself, one can understand why self-deception does not require that knowing and believing opposite things is simultaneous or that both believing and knowing are conscious and explicit. When intending to make oneself believe what one knows is false, one intends either to acquire the preferred belief or to get rid of one's knowledge, and in either case one is not interested in preserving one's conscious and explicit knowledge. The self-deceiver knows some proposition *p* to be false, and he or she makes himself or herself believe that *p* is true.

Still, the condition of knowing that *not-p* has its implications with regard to deliberate self-deception. Since one cannot consciously and explicitly believe something that one consciously and explicitly knows to be false, knowledge of the falsity of the proposition *p* must be somehow undermined. As long as I consciously and explicitly know that *p* is false, all my attempts to make myself believe that *p* will fail. The question 'How is

it possible to undermine one's own knowledge?' seems to be a hard one, unless one answers that it is impossible to undermine one's knowledge. Certainly, it is very hard to imagine that one could cease to know, for example, one's own age or the number of legs a butterfly has.

3.4.2 To Forget What One Knows

While it is true that one usually would know one's age, it is also true that one does not know, or is not aware of, some things one knew some time ago. For example, I do not know the basic laws of thermodynamics that I knew ten years ago. Naturally, one way one could seek to undermine one's knowledge is to try to forget something one knows. I would like to know whether I could intend to forget something I know and realize this intention.

The most straightforward approach is suggested by Thomas Schelling. According to him, there are many things one can do with one's mind, and he enumerates a list of different things that people use in order to forget what they want to forget (185-190). For example, one can sleep so that one would not think about unpleasant things, one can use alcohol or watch movies. Schelling himself suggests that self-deceivers could pick up something from this menu (184). I am afraid this menu does not suit my interests at this moment. The methods Schelling suggests remind me of the example of the drug dealer who, in some sense, deceived himself by undergoing a seance of hypnosis (see Section 2.2.-1). It is possible that one can get rid of one's knowledge using alcohol, hypnosis or sorcery, but one would not call such a manipulation of one's beliefs self-deception or deliberate self-deception. Self-deception seems to require that one succeeds in deceiving oneself using only one's mental capacities.

The easiest thing one could propose is to avoid thinking about something one knows

and to hope that the knowledge will disappear by itself. Any attempts to forget in this way, for example, the name of the capital of Canada would be a hopeless enterprise. But, certainly, one could hope that such a method will help one forget certain facts about some period in one's life and make it look better in one's own eyes. It is hard to say what precisely is the difference between one's knowing the capital of Canada and one's knowing some episode from childhood, but there must be some difference, if it is easier to forget one than another. Probably, the fact about the capital of Canada is more useful in everyday life than childhood memories, and there are many more occasions when one is reminded of the name of the capital than there are occasions when one is reminded of childhood memories. In any case, I would like to know how I could forget these memories.

Psychologists Daniel M. Wegner and David J. Schneider also have asked the question whether it is possible to forget what one knows. Their background interest seems in some way related to mine. They want to explore the possibilities of "psychological self-help" that can be understood as the management of unwanted thoughts and "the unwanted realities that those thoughts represent" (300). They suggest that the method by which such suppression can be done is by concentrating one's attention on something other than the thought that one wants to suppress.

The idea that the directing of one's attention could help one to deceive oneself is mentioned by several philosophers. Mele suggests that self-deceivers can intentionally shift their attention from unpleasant thoughts (*Irrationality* 126). Baghramian lists several 'strategies' self-deceivers can use, and one of the strategies suggests a shift of attention. Baghramian characterizes the strategy as avoiding "the undesirable thoughts or conclusions by keeping one's mind occupied with unrelated matters" (91). Davidson thinks that self-deceivers can intentionally direct their attention from the evidence that favours some undesirable belief *p* and so cause themselves believe the negation of the *p* ("Deception" 88).

It is certainly true that sometimes one intentionally shifts one's attention from one thought to another, and the thought one entertained first can disappear for good. For

example, I remember something unpleasant Mary told me the other day, and I immediately, without having a second thought about her words, turn my attention towards the problem of self-deception. Every time I remember what she said, or part of it, I think about something else. It is quite possible that after some time I will not be able to remember what exactly she said, while most likely I will remember that she said something unpleasant.

In the case I just described the shift in attention was not intended beforehand. Whenever the thought about Mary's words came up in my mind, I shifted my attention to something else and did it quite spontaneously. I shifted my attention not because I intended to do so, but because the words she said were unpleasant and I did not want to think about them. However, what I want to know is whether I can get rid of my memory of her words, if I intended to get rid of it.

Wegner and Schneider have tried to perform an experiment where people are asked not to think about a white bear. Such a request is certainly a request to shift one's attention deliberately away from some thought, or in other words, to deliberately forget something about which one is thinking. According to them, this experiment demonstrates whether one can control one's memories using what they call 'primary suppression' and 'auxiliary concentration.' Auxiliary concentration is "attending to something because we wish to suppress attention to something else;" primary suppression is "keeping attention away from something because we want to do so" (290). I will not present all the details of the experiment. Basically, the subjects of the experiment are asked to think out loud while trying not to think about a white bear. Each time they think about one, they have to ring a bell (296). The results of the experiment are clear -- the subjects are not able to suppress effectively their thoughts about the white bear, while they succeed better when they focus their attention on one particular thing and not just try to think about anything that is not a white bear (297-299). The final recommendation Wegner and Schneider can give to somebody who wants to get rid of some thought is "to avoid suppression, to stop stopping [one's thoughts]" (300). Such a recommendation must be understood as admitting of the impossibility of

suppressing some thought deliberately and, consequently, admitting the impossibility of making oneself forget something one knows.

Wegner and Schneider themselves have indicated that one must be cautious about generalizing the experiment's results (301). According to them, the thought about a white bear is different from the thoughts one tries to suppress in everyday life; the latter are usually charged with different emotional attitudes. Moreover, the requirement that the subjects must report their thoughts aloud adds some artificiality to the situation. I must agree that the thoughts one usually tries to suppress have emotional aspect, but I would also add that thoughts that are charged with emotions are usually harder to suppress than those that are not. When there is some thought that really bothers me, I cannot get rid of it, while if I think about grasshoppers, I can quite easily find something more interesting to think about and forget about grasshoppers. As Kierkegaard writes in *Either/Or*, the ability to forget depends on how one remembers things and, ultimately, how one experiences things (293).

In order to forget easily, one has to experience everything without being amazed, without enjoying anything too much or yielding to pain, i.e., without emotion. Nevertheless, emotional neutrality is not enough for suppressing a thought; after all, the subjects of the experiment could not suppress a thought that was emotionally neutral. I think that the problem lies not in the fact that the subjects of the experiment were asked to think aloud. I would rather think that the whole situation of experiment makes forgetting the white bear impossible, because one cannot forget why one is in the situation where one has to suppress the thoughts about the white bear. Meanwhile, the situations of one's everyday life need not remind one of the thought one tries to suppress. In addition, the situation of the experiment must be interesting enough in itself, and that can make any shift in attention very hard. So, it seems to me that the best circumstances for intended suppression of a thought must be such that the thought does not have any emotional importance, it does not have any practical importance, such as one's knowledge of the name of the capital of

Canada, and preferably, the circumstances are such that there are more interesting thoughts to think about than the thought one wants to suppress.

Of course, one has not ceased to know that *p*, if one has succeeded several times to suppress the thought that *p*. There is always a possibility that the thought will appear again, and it seems that there is no way to secure the state of suppression except by shifting one's attention from the thought again and again. It is possible that the thought will never come to one's mind again, but one can never know when the knowledge that *p* is undermined so that it is never expressed in one's awareness again. And there is no method that guarantees that the final suppression of knowledge will ever happen. The absence of means to achieve the suppression of knowledge could suggest that this suppression is not intended.

Elster, for example, thinks that the state of forgetfulness is a by-product of actions undertaken for ends other than forgetting something (*Sour* 48). For example, the by-product of my reading a book is the fact that I forgot to call Mary. According to Elster, one cannot intend to achieve the desirable state of forgetting. I cannot intend to forget to call Mary and intend to do that by reading a book. The state comes, if it comes at all, as a supplement, or by-product of some other action. At the same time, he does not deny that one could acquire the state of forgetfulness, but he insists that one should distinguish between the outcome of an action that is foreseen and that is intended (*Sour* 55). According to him, the states that are by-products cannot be intended, but only foreseen. I think that I must disagree with this analysis. The intended, or deliberate, action of making oneself forget what one knows is based on a principle that one starts one's non-basic action by shifting one's attention away from the thoughts that *p* and hopes that at one moment the thought and the knowledge that *p* will disappear. Of course, the final phase of such action is not directly controlled by the agent, but the same can be said about many other intended actions. For example, when Jones throws a ball into a basketball net, Jones directly controlled the flight of the ball when he threw it, but afterwards there was no possibility to control the flight, and certainly the ball could have missed the net. I would certainly call

this action an intended action, and I would not say that the ball's falling into the net is a by-product of the action. While Jones could not fully determine whether the ball would fall into the net, he wanted and intended it to fall there; and I think that Jones' wanting and intending suggest that the ball's falling into the net was not a by-product of the action of throwing the ball. Similarly, I think that one can call intended action one's making oneself forget what one knows, despite the fact that one cannot directly control the forgetting of the proposition *p,* or one's knowledge that p is false. I also think that when Davidson claims that self-deception must be intended and self-deceivers must intentionally direct their attention away from the important evidence, he has to accept this interpretation of forgetting.

I tried to show that there are good reasons to believe that one can intend to forget something and realize one's intention by diverting attention away from the thought, proposition, or belief one wants to forget. Still, even if one can make oneself forget what one knows, I have not shown that one can make oneself believe what one knows is false. There is no guarantee that when I forget that butterflies have six legs, I will believe that butterflies do not have six legs, or that they have eight legs and are in fact spiders. Most likely, if somebody suggested that butterflies have eight legs, I would remember the fact that I had successfully forgotten about butterflies having six legs. It seems that in order to make oneself believe what one knows is false, one has to undermine one's knowledge so that believing the opposite to what one knows is possible.

3.4.3 Reinterpretation of Evidence

I think that the most plausible method which one could use to undermine one's knowledge is the reinterpretation of the evidence for one's beliefs. Beliefs that I can be absolutely sure about are just a small fraction of my beliefs, and a bit of uncertainty is already a possibility for undermining the belief. Of course, many beliefs that I have seem

to me justified and true. I do not doubt that I am sitting now at my table. I do not doubt that I ate my breakfast this morning and have not had my lunch yet. Nevertheless, a rigorous Sceptic or a wise Buddhist could challenge these certainties. I do not really doubt that the Moon travels around the Earth, that the Europeans of the Middle Ages did know potatoes, or that in the Permian Period dragonflies were up to 70 cm long. Still, there is a chance that some great Scientist will come along, and these beliefs will turn out to be wrong. Thinking about the possibility of deliberate self-deception, I want to mention other beliefs that have their own certainties and uncertainties. If deliberate self-deception is possible, one must look at the beliefs that can be doubted without involving a group of nuclear physicists or specialists in 18th century art. Persons who are called self-deceivers usually would be self-deceived about some everyday things and problems, and I think that, first of all, I should look for deliberate self-deception among beliefs that concern problems of everyday life.

An interesting aspect of self-deception is the fact that sometimes argumentation against the self-deceiver's beliefs fails to convince him or her. I can present evidence and arguments and be sure that the evidence and arguments I present are overwhelming and justify my (the "correct") belief beyond any reasonable doubt, and when the self-deceiver still does not want to accept my claim or clings to his or her own, I decide that he or she is irrational, stupid or pretending not to understand my argument. It seems to me that there is a better explanation for such reluctance to accept reasons that persons who are not deceiving themselves would accept without doubt. I think that what, in fact, I have to look for is another possible interpretation of the evidence that for me seems to point in one direction. I must try to detect what makes the self-deceiver's interpretation possible.

Let me look at some examples that philosophers have used when they talk about self-deception. Amelie Oksenberg Rorty in her article "The Deceptive Self: Liars, Layers, and Lairs" describes the awkward and enigmatic behaviour of Dr. Laetitia Androvna (11). Dr. Laetitia Androvna is a specialist in the diagnosis of cancer. Usually she is perceptive

and does not avoid open discussion with her friends. Unfortunately, it seems that she has a cancer. The awkward thing about Dr. Laetitia Androvna is that she does not recognize the symptoms of her cancer, while they are so obvious that anybody who has the slightest knowledge of medicine would recognize them. Moreover, the doctor "uncharacteristically deflects their [friends of hers] questions and attempts to discuss her condition" (11). At the same time, Dr. Laetitia Androvna is drawing up a will and writes letters to friends and relatives. How can Laetitia Androvna be so inconsistent and seem not to recognize that she is such?

Of course, it is hard to know what exactly happens in Laetitia Androvna's mind, but I certainly could try to order all the facts in a manner that could give an explanation of her behaviour. First of all, to have a cancer is not quite the same as to have a wooden leg -- it is possible for both Laetitia Androvna and her friends to be mistaken about the nature of the trouble. If Laetitia Androvna avoids going to a doctor, she probably can find other explanations for her symptoms. She may be aware that there is a possibility that she have a cancer, but since she is not visiting a physician and is not talking to her friends about the problem, she can keep herself in the uncertain state that she prefers to the knowledge about the state of her health. While there is some possibility of interpreting her symptoms as being symptoms of something other than cancer, she entertains the thought that the symptoms will disappear and that she does not have cancer. The fact that she writes a will is nothing surprising. If she is aware that she could have cancer, she probably would consider it wise to write a will, just like one would leave home with an umbrella, if the forecast suggest that it could rain.

The most popular example among philosophers is the example of adultery. Using Siegler's example, I could tell the story about Brown's wife who is obviously unfaithful to Brown, but Brown believes that she is not (473). While to anybody else it seems obvious that the wife is unfaithful, there are also obvious opportunities to explain her behaviour without mentioning unfaithfulness. Unless Brown has witnessed a wild orgy involving his

wife and her lover, it is possible to imagine the interpretation of evidence as evidence not for adultery but for something else. For example, coming home later than usual need not necessarily mean that one is spending the extra time in a restaurant with the lover. People often have to work later than usual, and why should the husband think about adultery as the first possible explanation?

Jeffrey Foss gives an example of a mother who convinces herself that her son will not be paralysed even though the medical testimony indicates that he will (242). Since it is possible to imagine a situation when the paralysis is averted despite the bad condition of the patient, the mother is able to convince herself that her son will escape paralysis. A man, mentioned by Demos, convinced himself that he was a great womanizer and that "he has had interesting adventures with the ladies" (591). The notion of 'interesting adventures' has no strict meaning, and only in exceptional cases would one be unable to find anything that confirmed this perception of oneself.

Certainly, it is possible to explain things in a different way. One can even notice certain areas where it is easy to find some justification for false beliefs. Very rarely philosophers would talk about self-deception that concerns something that one can see with his own eyes or hear with his own ears. The only exceptions are Sackeim and Gur who ascribe self-deception to persons who do not recognize consciously their own voice when it was played to them (173-175). Nevertheless, usually when self-deceivers deceive themselves about some present situation, the evidence for their beliefs is usually indirect. If a husband wants to explain his wife's returning from work late, he must explain this using his knowledge of his wife's character or her past.

The interpretation of the past, one's own or somebody else's, provides a great opportunity for a justification of false beliefs that one wants to believe. One example could be Demos' 'womanizer.' And such examples could be many. For example, if Jones sees within himself a great leadership talent and his belief is challenged, Jones can find in his past something that would somehow justify his belief that he has leadership talents. If

Jones avoids testing his talents in practice, he can sustain his belief in his talents for very long time, however insignificant the evidence for this belief could be.

My future also can be interpreted according to my interests. Even if evidence is against my belief about something that I expect to happen in the future, I can be confident that my own petty *deus ex machina* will emerge from nothingness and rearrange things so that they fit my expectations. A classical example is Hitler's belief in the victory of German troops despite the fact that the Allies were already in Germany.

The possibility of interpreting evidence according to one's preferences can certainly shed some light on unintentional self-deception. There is still the question whether one can deceive oneself deliberately. The hard thing about deliberate self-deception is the condition that one knows something and tries to make oneself believe the opposite. A simple interpretation of the evidence is not useful for deliberate self-deception because when the person knows something he or she has already some interpretation of the evidence and, according to my definition of deliberate self-deception, this interpretation is a correct one. If deliberate self-deception is possible one must be able to interpret the evidence so that the correct justification would look to the self-deceiver to be incorrect, that is to say, one must be able to reinterpret the evidence.

At first, such a project may look easy. As I showed, in certain circumstance the evidence can favour a preferable, but a wrong belief. If one can interpret evidence in a certain way, one should have been able also to reinterpret it. Nevertheless, there are two difficulties: one conceptual and one practical. The conceptual difficulty is such that the condition of knowledge seems to presuppose that the belief is justified so that there cannot be any doubt about its correctness. If there is no uncertainty about the truth of the belief and the way the evidence should be interpreted, no reinterpretation is possible and knowledge cannot be undermined. The second difficulty is such that since self-deception is deliberate, the self-deceiver knows that he or she is biasing the evidence and this knowledge can undermine the whole project of self-deception.

Of course, if knowing is understood in the strictest sense, no deliberate self-deception is possible. Nevertheless, usually one's true beliefs are not justified to such an extent that no other interpretation could be possible. In everyday practice, beliefs are justified only reasonably well. I know that the Post Office is open today, since I have been there. I know the closing hours of the Post Office and I have not heard that anything bad has happened there. I can say that I know that the Post Office is open and my belief is justified. It is even possible that my belief is true. Nevertheless, nothing can prevent me from entertaining a plausible thought that it is closed right now. I remember that the employee at the Post Office looked a little bit sick, so it is quite plausible that he felt so bad that he went home and the Post Office is closed now. So if I had to but did not want to go to the Post Office, I had a reason for postponing my going there. After all, the Post Office could be closed and the walk would be futile.

As one can imagine, by the previous line of reasoning I did not convince myself that the Post Office is closed. I do not really believe that it is closed now, because I know that I invented the sick employee. But could I deceive myself if the post office employee looked sick? It is hard to answer the concrete example about the employee at the Post Office, but it seems to me that under certain circumstances I could have found an interpretation of evidence that favoured this belief. For example, I could remember different stages in the writing of this thesis and come up with different stories of how I wrote it: 'I really did not work hard for several months, maybe only at the end;' 'That was horrible, I do not understand how I got to the end;' 'From the beginning I had the plan and the main ideas of the thesis in my mind, so I just had to put everything on the paper.' I think deliberate self-deception is possible.

3.5 Summary

When I chose to write my thesis about the possibility of deliberate self-deception, I

did not assume that such self-deception is possible. One need not read many books in order to know that one cannot make oneself believe just whatever one wants to believe. At the same time, I did not assume that deliberate self-deception is never possible, and I wanted to know whether any philosopher has provided some clue as to how one could deliberately deceive oneself.

I must say the strategies philosophers ascribe to self-deceivers do not contain any surprising ideas on how one can make oneself believe something one knows or believes to be false. Still, I tried to show that purposeful reinterpretation of evidence for one's beliefs can form what I call deliberate self-deception. The problem with this self-deception is that the evidence can be reinterpreted so that one's knowledge is undermined only in some cases. Deliberate self-deception cannot be realized on many occasion when one perhaps would like to make oneself believe what one knows is false. Though I claimed that one can deliberately reinterpret evidence for one's beliefs and also that in certain circumstances one can forget what one knows, I must agree with Elster that all these methods of controlling one's mind are "too costly" (*Sour* 57). With the phrase 'too costly,' he certainly does not want to suggest any monetary expenses. He claims that an intended bringing about of a mental state could be technically possible, but usually the sacrifices one must make in order to discipline one's mind outweigh the benefits received from the desired mental state. One can try to make oneself believe that the Post Office is closed, but usually one would not bother to persuade oneself to believe anything so trivial. Still, I think that this analysis of deliberate self-deception allowed me to look at different aspects of self-deception and the possibility to control one's beliefs.

A large part of the thesis concerned the understanding of self-deception in ordinary language and philosophical discourse. I think that the analysis of the term 'self-deception' and how this term is used by philosophers helped me to clarify many aspects of the discussion of self-deception. Furthermore, I tried to contribute something to a better understanding of self-deception and human mentality in general.

If I had to evaluate which aspects of self-deception seem to be the most interesting ones for further studies, I would mention two of them. The first is directly connected with the problem of deliberate self-deception. Several philosophers have suggested that there are certain aspect of intentionality and purposefulness in self-deception. As my analysis shows, the possibility to deceive oneself deliberately are quite scarce. There are several problems that would be interesting to analyse. For instance, what is the difference between intentional and deliberate deception that allows the former to be realized easier than the latter? Why are certain actions are not successful when intended beforehand? Why does one's awareness that the evidence is selected intentionally undermine the self-deception? Why is a self-deceiver who does not deceive himself or herself deliberately not aware that the evidence one has is selected? In a word, what is the function of awareness in self-deception?

The second aspect of self-deception that is worth a closer look is the role of language in self-deception. The aim of my thesis was to analyse the possibility of deceiving oneself deliberately, and I was not able to look closer at the way people present their beliefs, experiences and interpretations of different aspects of their life. The interpretations of evidence and experiences usually are expressed in language and have the form of a narrative. It is very probable that one's preconceptions, learned and traditional interpretations of the self, others and one's environment may shape one's actual experience and result in biasing of beliefs.

LITERATURE CITED

Anscombe, G. E. M. *Intention*. Oxford: Blackwell, 1957.

Audi, Robert. "Self-Deception and Rationality." *Self-Deception and Self-Understanding*. Ed. Mike W. Martin. Lawrence, Kansas: University Press of Kansas, 1985. 169-194.

Baghramian, Maria. "The Paradoxes of Self-Deception: A Replay to R. T. Allen." *Irish Philosophical Journal* 7.1-2 (1990): 171-179.

---. "Strategies of Self-Deception." *Irish Philosophical Journal* 3.2 (1986): 83-97.

Barnes, Anette. "When Do We Deceive Others?" *Analysis* 50.3 (1990): 197-202.

Bratman, Michael E. "Intention." *A Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. Oxford: Blackwell, 1994. 375-379.

Butler, Joseph. "Upon Self-Deceit." *Vice and Virtue in Everyday Life*. Ed. Christina Hoff Sommers. San Diego: Harcourt Brace Jovanovich, 1985. 264-270.

Canfield, John V., and Don F. Gustavson. "Self-Deception." *Analysis* 23 (1962): 32-36.

Champlin, T. S. "Deceit, Deception and the Self-Deceiver." *Philosophical Investigations* 17.1 (1994): 53-58.

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

Danto, Arthur. "Basic Actions." *American Philosophical Quarterly* 2.2 (1965): 141-148.

Davidson, Donald. "Deception and Division." *The Multiple Self*. Ed. Jon Elster. Cambridge: Cambridge University Press, 1985. 79-92.

---. "Intending." In *Essays on Actions and Events*. Oxford: Clarendon, 1980. 83-102.

Davis, Lawrence H. "Action." *A Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. London: Blackwell, 1995. 111-117.

---. *Theory of Action*. Englewood Cliffs, New Jersey: Prentice-Hall, 1979.

"Deliberate." *The Oxford English Dictionary*. 1989 ed.

Demos, Raphael. "Lying to Oneself." *Journal of Philosophy* 57 (1960): 588-595.

Dyke, Daniel. *The Mystery of Selfe-Deceiuing*. London: William Stansby, 1633.

Elster, Jon. *Sour Grapes*. Cambridge: Cambridge University Press, 1983.

---. *Ulysses and the Sirens*. Cambridge: Cambridge University Press, 1979.

Fingarette, Herbert. *Self-Deception*. London: Routledge & Kegan Paul, 1969.

---. "Self-Deception and the 'Splitting of the Ego.'" *Freud: A Collection of Critical Essays*.

Ed. Richard Wollheim. Garden City, New York: Anchor Books, 1974.

Foss, Jeffrey. "Rethinking Self-Deception." *American Philosophical Quarterly* 17.3 (1980): 237-243.

Gardiner, Patrick. "Error, Faith and Self-Deception." *Proceedings of the Aristotelian Society* LXX (1970): 221-244.

Goldman, Alvin I. *A Theory of Human Action*. Englewood Cliffs, New Jersey: Prentice-Hall, 1970.

Guttenplan, Samuel. "Self-Deception." *A Companion to the Philosophy of Mind*. Ed. Samuel Guttenplan. London: Blackwell, 1995. 558-560.

Haight, Mary R. *A Study of Self-Deception*. Brighton, Sussex: The Harvester, 1980.

---. "Tales from a Black Box." *Self-Deception and Self-Understanding*. Ed. Mike W. Martin. Lawrence, Kansas: University Press of Kansas, 1985. 244-260.

James, William. "The Will to Believe." In *"The Will to Believe" and Other Essays in Popular Philosophy*. New York: Dover, 1956. 1-31.

Johnson, Samuel. "No. 28. Saturday, 23 June 1750." In *The Rambler*. Vol. III of *TheYale Edition of the Works of Samuel Johnson*. Ed. W. J. Bate and Albrecht B. Strauss. New Haven and London: Yale University Press, 1965. 151-157.

Kearney, Richard. *Dialogues with Contemporary Continental Thinkers*. Manchester: Manchester University Press, 1984.

Kierkegaard, Soren. *Either/Or*. Ed. and trans. Howard V. Hong and Edna H. Hong. Princeton, New Jersey: Princeton University Press, 1987.

King-Farlow, John. "Self-Deceivers and Sartrian Seducers." *Analysis* 23 (1963: 131-136.

Kipp, David. "Self-Deception, Inauthenticity, and Weakness of Will." *Self-Deception and Self-Understanding*. Ed. Mike W. Martin. Lawrence, Kansas: University Press of Kansas, 1985. 261-283.

Martin, Mike W. "General Introduction." *Self-Deception and Self-Understanding*. Ed. Mike W. Martin. Lawrence, Kansas: University Press of Kansas, 1985. 1-28.

McLaughlin, Brian P. "Exploring the Possibility of Self- Deception in Belief." *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie Oksenberg Rorty. Berkeley: University of California Press, 1988. 29-62.

Mele, Alfred. *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press, 1987.

---. "Recent Work on Self-Deception." *American Philosophical Quarterly* 24.1 (1987): 1-17.

Paluch, Stanley. "Self-Deception." *Inquiry* 10 (1967): 268- 278.

Pears, David. "Freud, Sartre and Self-Deception." *Freud: A Collection of Critical Essays*. Ed. Richard Wollheim. Garden City, New York: Anchor Books, 1974. 97-112.

---. "The Goals and Strategies of Self-Deception." *The Multiple Self*. Ed. Jon Elster. Cambridge: Cambridge University Press, 1985. 59-78.

Penelhum, Terence. "Pleasure and Falsity." *American Philosophical Quarterly* 1.2 (1964): 81-91.

Plato. *Cratylus*. Trans. Benjamin Jowett. *The Collected Dialogues of Plato*. Ed. Edith Hamilton and Huntington Cairns. Princeton, New Jersey: Princeton University Press, 1961. 421-474.

Price, H. H. "Belief and Will." *Aristotelian Society*, Supplementary Vol. XXVIII (1954): 1-26.

Radden, Jennifer. "Defining Self-Deception." *Dialogue* XXIII (1984): 103-120.

Raleigh, Walter. *Shakespeare*. London: Macmillan 1907.

Rorty, Amelie Oksenberg. "The Deceptive Self: Liars, Layers, and Lairs." *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie Oksenberg Rorty. Berkeley: University of California Press, 1988. 11-28.

---. "Self-Deception, *Akrasia* and Irrationality." *The Multiple Self*. Ed. Jon Elster. Cambridge: Cambridge University Press, 1985. 115-132.

Ruddick, William. "Social Self-Deception." *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie Oksenberg Rorty. Berkeley: University of California Press, 1988. 380-389.

Sackeim, Harold A., and Ruben C. Gur. "Self-Deception, Self-Confrontation, and Consciousness." *Consciousness and Self-Regulation.* Vol. 2. Ed. Gary E. Schwartz and David Shapiro. New York: Plenum, 1978. 139-197.

Shaffer, Jerome A. *Philosophy of Mind*. Englewood Cliffs, New Jersey: Prentice-Hall, 1968.

Schelling, Thomas. "The Mind as a Consuming Organ." *The Multiple Self*. Ed. Jon Elster. Cambridge: Cambridge University Press, 1985. 177-196.

Schmitt, Frederick F. "Epistemic Dimensions of Self-Deception." *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie Oksenberg Rorty. Berkeley: University of California Press, 1988. 183-204.

Scott-Taggart, M. J. "Socratic Irony and Self Deceit." *Ratio* 14 (1972). 1-15.

Searle, John. *Intentionality*. Cambridge: Cambridge University Press, 1983.

"Self-Deception." *The Oxford English Dictionary*. 1989 ed.

Siegler, Frederick A. "Demos on Lying to Oneself." *The Journal of Philosophy* 59 (1962): 469-475.

Silver, Maury, John Sabini and Maria Miceli. "On Knowing Self-Deception." *Journal of the Theory of Social Behaviour* 19.2 (1989): 213-227.

Talbott, W. J. "Intentional Self-Deception in a Single Coherent Self." *Philosophy and Phenomenological Research* 55.1 (1995): 27-74.

Van Fraassen, Bas C. "The Peculiar Effects of Love and Desire." *Perspectives on Self-Deception*. Ed. Brian P. McLaughlin and Amelie Oksenberg Rorty. Berkeley: University of California Press, 1988. 123-156.

Wegner, Daniel M. and David J. Schneider. "Mental Control: The War of the Ghosts in the Machine." *Unintended Thought.* Ed. James S. Uleman and John A. Bargh. New York:

Guilford Press, 1989. 287-305.

Williams, Bernard. "Deciding to believe." In *Problems of the Self: Philosophical Papers 1956-1972*. Cambridge: Cambridge University Press, 1973. 136-151.