SELF-DECEPTION

A Critique of Alfred Mele's Theory of Self-Deception

by

YIXI ZHANG

B. A., The University of British Columbia, 1995

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Department of Philosophy)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

April, 1999

© Yixi Zhang, 1999
In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Philosophy

The University of British Columbia
Vancouver, Canada

Date April 27, 1999
This thesis deals with some basic philosophical problems of self-deception. Following a detailed exposition of Alfred Mele's account of self-deception, it shows that Mele's account is inadequate but can be improved by a reason explanation of self-deception that is based on a concept of deception narrower than the one that Mele has been using.
# TABLE OF CONTENTS

Abstract ........................................................................................................... ii

Acknowledgements ......................................................................................... iv

Chapter 1 Overview and Introduction ............................................................ 1

Introduction ................................................................................................... 1
1. How to Look for Self-Deception ................................................................. 6
2. Problems of Self-Deception ....................................................................... 11
3. Desire and Motivation in the Theory of Self-Deception ......................... 15
4. Related Phenomena .................................................................................. 16

Chapter 2 Mele’s Theory of Self-Deception ................................................... 19

1. Two Paradoxes of Self-Deception .............................................................. 19
2. Self-Deception Described .......................................................................... 25
3. Entering Self-Deception ........................................................................... 28
4. The Strategies of Self-Deception ................................................................. 33
5. Reasons and Motivation .......................................................................... 40
6. Self-Control, Continent Action and Balanced Motivation ..................... 45
7. Conclusion ............................................................................................... 48

Chapter 3 A Critical Examination of Mele’s Theory ..................................... 50

1. Deception and Self-Deception .................................................................. 50
2. Desire ......................................................................................................... 64
3. Motivation ................................................................................................. 67
4. An Explanation of Self-Deception ............................................................... 86
5. Conclusion ................................................................................................ 89

Chapter 4 Related Phenomena and Conclusion ............................................ 91

1. Self-Induced Deception ........................................................................... 91
2. Wishful Thinking ..................................................................................... 93
3. Naivety and Ignorance ............................................................................ 95
4. Akratic Believing and Irrationality ............................................................ 98
5. Lying to Oneself ....................................................................................... 101
6. Pseudorationality ................................................................................... 102
7. Conclusion ............................................................................................... 104

Bibliography ................................................................................................. 107
ACKNOWLEDGMENTS

This thesis was made possible by the continuous encouragement and help given by various people. I am grateful to both Dr. Catherine Talmage and Dr. Andrew Irvine in the Department of Philosophy at the University of British Columbia for their suggestions, comments and meticulous editorial work on this thesis.

I wish to express my gratitude to my mentor, Dr. Catherine Talmage, for her strong and constant guidance to me in every turn of my academic and personal development. Over the past four years, Dr. Talmage has been much more than a supervisor to me. With her elegant frame of mind and penetrating way of thinking, she has taught me to be a better philosopher; with her warmth and deep honesty, she has cultivated me to be a better and happier person. Kate, I am forever indebted to you.

I am extremely grateful to another mentor of mine, Dr. Andrew Irvine, for being one of the most reliable academic supporters, ever since my junior undergraduate years, of my pursuit of philosophy and beyond. Dr. Irvine’s insight and sharpness of mind, rich academic experience, moral integrity and sensitivity to issues of gender and ethnicity have been and will always be a strong inspiration in my life. Andrew, you are my hero.

I am also indebted to my partner Daniel Bryant for his love, devotion and his ability to make me smile. He has not only carefully read through this thesis many times during different stages of its development and has made the most of the editorial comments and suggests, but also wiped the muddy paws of my puppy when I was away.
CHAPTER 1

Overview and Introduction

Everything about the concept of 'self-deception' is controversial among philosophers, beginning with its definition.

-- Sissela Bok

Introduction

Why is there self-deception? Why do so many otherwise intelligent people at times appear to be so oblivious, so blind, or so dumb? Why do people in abusive relationships often believe that everything is all right? Why did many Chinese intellectuals, for the three decades during Mao Tsetung's reign, sincerely believe in the Communist totalitarian government, despite overwhelming evidence that Mao was one of the cruelest rulers in Chinese history?

These questions seem to suggest that general intelligence and the availability of relevant evidence sometimes have little effect on belief acquisition or retention. I believe that the truth value of our beliefs ultimately depends on the world, and that the world, at times, forces its facts upon us with such power that we cannot but believe them. Nevertheless, we sometimes acquire or retain false beliefs as a result of our own acts. For instance, I can make beliefs about myself false by intentionally behaving in a certain way, so that my action makes the belief false. My belief that I will take a walk this afternoon can be made false by my staying inside the house all afternoon. Self-fulfilling prophecy also explains my point. However, this type of belief manipulation is not of present interest. What I am interested in is the fact that, as rational creatures, motivated by desire, fear or certain beliefs, we treat ourselves in such a way that we end up believing something false.

Ideally, the job for our mind is to fit its beliefs to the way the world is. Usually, we

1Bok, 1995, pp.818-819.
believe that $p$ because we believe that it corresponds to the fact that $p$. However, in addition to beliefs, we also have wishes, fears and desires. We theorize and give reasons for what we believe, and what we ought to believe, including what we want to believe, even though we ought not to. The forming or retaining of a belief is therefore not independent of the existence of all the other attitudes and theories that we have. There are conflicting moments during which we would compromise the truth for ease of mind. We would manage to believe something for reasons other than that it is the case. We would, as it were, by wanting the world to fit our belief system, end up having false beliefs about ourselves and the world. Self-deception is one such case.

Self-deception is a multi-disciplinary subject. It has its moral dimension. Questions such as ‘Does self-deception bring moral good while not being epistemically justifiable?’ or ‘Is self-deception as much an ethical issue as an epistemic one?’ are often raised. Kant argues that “the formal condition of all virtues” is sincerity with self. Bishop Butler points out that there is a danger of having the “deep and calm source of delusion” affect us. That is, one can be self-deceived in believing that one is moral by acting according to certain moral rules while in fact one is being less than moral. Being self-righteous is an example of this. A self-righteous person is, according to Stephen Darwall, self-deceived in believing that he is a morally good person but, instead of being concerned about acting in morally appropriate ways, the self-righteous person’s concern is to think of himself as worthy of approval. A person may self-regulate in order to be a moral person and conceive of himself as a morally good person whose acts are always guided by his better moral judgment. The self-deception appears when, aiming

---

2Kant, 1964, p.103.

3Butler, 1900, pp.135-136.
at his moral ideal, he settles for his conception of the ideal instead of the reality. Moreover, he can be self-deceived into believing that he acts morally by rationalizing his act. Because his moral rationalization can ruin his sense of moral judgment, when he morally rationalizes his act, he is not only not acting morally but also preventing himself from improving as a moral person.\(^4\)

Self-deception has its social and political dimension. Self-deception may be thought of as some thing one does to oneself, but some philosophers argue that self-deception is not as private as it appears to be. The acquisition and retention of a self-deceived belief often needs the assistance of others. William Ruddick points out that a person's self-deception is often helped by people around him.

We ... choose the company of those whose views coincide with our own. Hence, our projects come to be questionable only from a perspective we are unlikely or even unable to take ... Even if we can 'open our moral eyes,' it is unlikely that we will be able to see very much: our associates, out of sympathy or cowardice, tend to keep the lights turned down low.\(^5\)

Amélie Rorty argues that socially induced self-deception promotes social cooperation and cohesion, even though the responsibility for self-deception lies primarily with the individual. Moreover, the danger of self-deception lies, not so much in the irrationality of the occasion, as in the consequence of the habit, since self-deception does not monitor its own use. Thus, Rorty cautions us to be careful about the company we keep.\(^6\)

Self-deception also has its cross-cultural dimension, which has become a relatively new area of study. It has been argued by some philosophers that the most fundamental elements of

---

\(^4\)Darwall, 1988, p.426.


\(^6\)Rorty, 1996, p.82.
self-deception, such as the nature of belief, rationality, and the concept of 'self', are differently conceived in different cultures. In addition, some of them raise the question of whether the phenomenon of self-deception is a cross-cultural issue. Others believe that self-deception varies its form from culture to culture. For instance, Roger Ames maintains that the Confucian 'self' is constructed as a non-unitary abstraction that shifts according to the social roles one plays, and that self-deception may be described and understood differently in traditional Chinese culture.\(^7\)

Self-deception has its psychological dimension, in which cognitive scientists have great interest.\(^8\) Social psychologists, for instance, have made many findings on the formation and maintenance of erroneous beliefs that are causally related to social interactions. One of the problems is what social psychologists call the 'false consensus effect', namely, the tendency for our own beliefs, values, and habits to bias our estimates of how widely such views and habits are shared by others. Another problem is that we usually do not get adequate critical feedback from others, either because we are in some way socially isolated or because people are usually reluctant to give negative feedback.\(^9\) Other problems, such as our tendency to play 'hot hand',\(^10\) which results in the misinterpretation of incomplete and unrepresentative data, can result in the biassed interpretation of data due to erroneous theories or preconceptions. Our tendency to

---

\(^7\) Ames, 1996, p.236.


\(^10\) This is the belief that is commonly held by basketball players that, after making a few good shots at the beginning of a game, the player is more likely to make subsequent shots. This idea is similar to the one that "success breeds success" and "failure breeds failure." Although this idea may be true in some cases, such as in financial success, it is erroneous when applied to situations such as sports. For more discussion on this topic, see Gilovich, 1991.
superimpose order on random, chaotic and illusory phenomena can contribute to self-deception.

Important and interesting as the above dimensions of self-deception are, they are beyond the scope of my present work. As a philosophy student, I think that it is appropriate to concentrate on what I have been trained to do and abstain from psychological discussions or cross-cultural studies. As I understand it, the philosophical study of self-deception has been, at least for the past two decades, largely concentrated on the analysis of the phenomenon and on its epistemic dimensions, as a part of philosophy of mind and philosophy of action. I will focus my discussion on the analysis and the epistemic dimension of the phenomenon in general, and on motivated irrationality in particular. I will explore the concepts of deception and self-deception, the nature of desire and motivation, the nature of irrationality and phenomena related to self-deception. The kind of self-deception that I am interested in is the sort that we run into often in everyday life among ordinary rational people. I will treat the self-deceived agent as a unitary self who is rational and is aware of his own beliefs and desires. My chief aim is to provide for a theory that provides a best explanation for such a phenomenon.

I lay out the basic problems and issues concerning self-deception in the present chapter. Chapter 2 is devoted to a discussion of Alfred Mele’s account of self-deception. Chapter 3 consists of my comments and criticisms of Mele’s account. At the end of Chapter 3, I will present my suggestion for characterizing self-deception. I conclude my thesis by discussing phenomena related to self-deception in Chapter 4.
1. How to Look for Self-Deception

Plato asks,

How will you look for it, ..., when you do not know at all what it is? How will you aim to search for something you do not know at all? If you should meet with it, how will you know that this is the thing that you did not know?\(^{11}\)

I do not think that we need to \textit{know} what self-deception is in order to look for it. Having some clear ideas about what self-deception might be will lead us a long way toward finding it. However, discovering the right answer to the question, "What is self-deception?" is one of the most difficult tasks in the study of self-deception. As we mentioned earlier, the phenomenon of self-deception is of interest to people from many different disciplines. When the question "What is self-deception?" is asked by philosophers, I take it to mean "What is, in a philosophically interesting sense, self-deception?" Narrowing the scope of self-deception to a philosophical level does not mean that the demarcation of the phenomenon is easy. An account of self-deception must first face the dilemma: On one hand, for the sake of the satisfaction of our philosophical intuition, we want to take the most subtle, and hence the most interesting, problems of the phenomenon into our account of self-deception. On the other hand, for the sake of theoretical elegance, or for the sake of the manageability of a complex phenomenon, we want to find the clearest and the simplest description we possibly can. Theories that focus on the former tend to lack analytical rigour, but those that focus on the latter tend to over-simplify the problem. The situation may be illustrated best by some examples given by various philosophers as cases of self-deception. I will first quote Sartre's example at length, and then I will discuss some other examples that are much simpler. In so doing, I hope to show that the

\(^{11}\)Plato, \textit{Meno}. 80d.
problems of self-deception can be as subtle as what Sartre describes, while the phenomenon may be elegantly explained by focussing on so-called “garden variety” cases that are much more manageable, though complicated in their own way.

My present work is largely a critique of Alfred Mele’s account of self-deception, which belongs to the simple and elegant type of work. This type of work is more manageable for my present purpose. However, I hope that, at the end of our discussion, readers will not go away with the belief that self-deception can be completely explained or can be explained away, regardless of how convincing some of the arguments may seem to be. Hence, it may be useful to have a look at a much more subtle case of self-deception before we launch into Mele’s simple and elegant analysis.

One of the most subtle cases of self-deception is Sartre’s story of a woman who went out with a man on a date. Sartre tells us that the woman knows the intention behind her suitor’s flattering behaviour and that she also knows that, sooner or later, she has to make a decision. But she does not want to realize the urgency ... She does not ... want to see possibilities of temporal development which his conduct presents ... If he says to her, “I find you so attractive!” she disarms this phrase of its sexual background; she attaches to the conversation and to the behaviour of the speaker, the immediate meanings, which she imagines as objective qualities. The man who is speaking to her appears to her sincere and respectful as the table is round or square, as the wall colouring is blue or gray. The qualities thus attached to the person she is listening to are in this way fixed in a permanence like that of things, which is none other than the projection of the strict present of the qualities into the temporal flux. This is because she does not quite know what she wants. She is profoundly aware of the desire which she inspires, but the desire cruel and naked would humiliate and horrify her. Yet she would find no charm in a respect which would be only respect. In order to satisfy her, there must be a feeling which is addressed wholly to her personality ... and which would be a recognition of her freedom. But at the same time this feeling must ... address itself to her body as object. This time then she refuses to apprehend the desire for what it is; she does not even give it a name; she recognizes it only to the extent that it transcends itself toward admiration, esteem, respect and that it is wholly absorbed in the more refined forms which it produces, to the extent of no longer figuring anymore as a sort of warmth and
density. But then suppose he takes her hand. This act of her companion risks changing the situation by calling for an immediate decision. To leave the hand there is to consent in herself to flirt, to engage herself. To withdraw it is to break the troubled and unstable harmony which gives the hour its charm. The aim is to postpone the moment of decision as long as possible. We know what happens next; the young woman leaves her hand there, but she does not notice that she is leaving it. She does not notice because it happens by chance that she is at this moment all intellect. She draws her companion up to the most lofty regions of sentimental speculation; she speaks of Life ..., she shows herself in her essential aspect -- a personality, a consciousness. And during this time the divorce of the body from the soul is accomplished; the hand rests inert between the warm hands of her companion -- neither consenting nor resisting -- a thing.

We shall say that this woman is in bad faith. But we see immediately that she uses various procedures in order to maintain herself in this bad faith. She has disarmed the actions of her companion by reducing them to being only what they are; ... But she permits herself to enjoy his desire, to the extent that she will apprehend it as not being what it is ... Finally while sensing profoundly the presence of her own body -- to the point of being aroused, perhaps -- she realizes herself as not being her own body, and she contemplates it as though from above as a passive object to which events can happen but which can neither provoke them nor avoid them because all its possibilities are outside of it. What unity do we find in these various aspects of bad faith? It is a certain art of form and idea and the negation of that idea ... The two aspects of human reality are and ought to be capable of a valid coordination. But bad faith does not wish either to coordinate them or to surmount them in a synthesis. Bad faith seeks to affirm their identity while preserving their difference in such a way that at the instant when a person apprehends the one, he can find himself abruptly faced with the other.12

This story is very richly told. Sartre’s analysis of the woman’s behaviour may be open to many different interpretations. Whether the woman in the story is self-deceived or not is highly controversial among philosophers.13 Nevertheless, by reading the story the way Sartre tells it, we may appreciate the sensitivity and the complexity of the issues related to self-deception. For one thing, the story suggests that self-deception is a kind of self-betrayal by virtue of being dishonest or inauthentic to one’s true being. One takes one’s genuine self to be

12Sartre, 1956, pp.96-98. (Italics mine).

something above and beyond the roles one plays, all the while being aware of, or taking
pleasure in, those roles. Whether or not such a way of treating oneself can bring pragmatic
good is a different issue. It may also be argued that displaying such behaviour occasionally may
not necessarily be a case of self-deception. Even so, to deny that the woman in the story is self-
deceived by arguing that she is only playing a fair game of dating misses the point. There is
no reason to think that game-playing will not lead us to self-deception. Even though lack of
authenticity itself may not be the same as self-deception, the frequent practice of it may easily
lead us into self-deception.

Self-deception can be viewed, in a doxastic sense, as one misleading oneself. The
garden-variety examples of self-deception provided by M. R. Haight may give us a general
picture of it: The cancer patient who ignores his symptoms or explains them away; the
drunkard who will not admit his alcoholism in the teeth of the evidence; the mother who is
blind to her son’s faults; the scientist who blandly denies that he shares any responsibility for
the consequences which may result from his research in biological warfare. Mele’s account of
self-deception will be aimed at these garden-variety cases of self-deception.

Having cited some cases of self-deception, now let us turn to the lexical approach
towards the phenomenon; i.e., let us define the term. Here again, different philosophers differ
in their definitions of self-deception. Davidson defines self-deception this way: \( S \) is self-
deceived with respect to a proposition \( p \) if and only if \( S \) has evidence for believing \( p \) but \( S \) has

\[\text{For further reading, see Kipp, 1985, pp.261-283.}\]

\[\text{For further reading, see Higgins,1996, pp.125-131.}\]

\[\text{Haight, 1980, p.82.}\]
motivation for believing not-\( p \), and \( S \) acts in such a way as to cause himself to believe not-\( p \).\(^{17}\) Alfred Mele has a very similar concept of self-deception, except that he insists that the self-deceived belief must be false. Robert Audi defines self-deception somewhat differently: \( S \) is self-deceived with respect to a proposition, \( p \), at time \( t \), if and only if, at \( t \), \( S \) unconsciously knows that not-\( p \) (or has reason to believe, and unconsciously and truly believes, that not-\( p \)); \( S \) sincerely avows, or is disposed to avow sincerely, that-\( p \); and \( S \) 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 \( S \) sees is evidence against \( p \).\(^{18}\) Amélie Rorty, on the other hand, argues that self-deception does not have to be directly motivated by the desire that \( p \). The self-deceiver who believes that \( p \) can be motivated by past desire that \( \neg p \). The student who falsely believes, right after the final examination, that she has failed her Philosophy 101 course may be motivated by the overwhelming desire to succeed that she had before the final.

Moreover, Rorty argues that self-deception need not involve false belief. A self-deceiver can not only get himself to believe something that in fact is true, but a self-deceiver may also be in a subdoxastic condition, hence have no belief at all with respect to something that he does not want to believe.\(^{19}\) I am sympathetic to Rorty's last point; however, the person who has a true belief that \( p \) or who does not have a belief that \( p \) cannot be said to be self-deceived into believing that \( p \); nor can he be said to be self-deceived in believing that \( p \). If one can establish an argument that the deceptive behaviour or deceptive disposition counts more than the truth of the belief when it comes to the attribution of deception, then one can say that an agent is

---

\(^{17}\)Davidson, 1985, p.144.

\(^{18}\)Davidson, 1985, p. 211.

\(^{19}\)Rorty, 1996, pp.77-78.
deceiving or self-deceiving whenever the deceptive behaviour is displayed, despite the fact that one holds no belief at all, or the belief held may in fact be true.

In short, self-deception can be conceived to be as complex as what Sartre has shown in his story, or as relatively simple and clear as the garden-variety cases we have presented.

2. Problems of Self-Deception

Many philosophers argue that self-deception, taken in the literal sense, is paradoxical, and that it is possible only in the metaphorical sense. Deception is often described in the following way: For any \( A \) and \( B \), when \( A \) deceives \( B \) into believing that \( p \), \( A \) knows or truly believes that \( \neg p \) while causing \( B \) to believe that \( p \). And self-deception may look like this: when \( A \) deceives \( A \) into believing that \( p \), \( A \) knows or truly believes that \( \neg p \) while causing himself to believe that \( p \). If the nature of one’s belief is such that one cannot believe both that \( p \) and that \( \neg p \), then self-deception seems impossible. Even so, self-deception does exist. So we have what Alfred Mele calls ‘the paradox of belief’. There is another paradox of self-deception, ‘the strategy paradox.’ It goes like this. In general, \( A \) cannot successfully employ a deceptive strategy against \( B \) if \( B \) recognizes \( A \)’s intention and plan. In cases of self-deception, \( A \) and \( B \) are the same person. Since one always knows one’s own intention and plan, one cannot successfully employ a deceptive strategy against oneself. Therefore, self-deception should be impossible. However, people are in fact often self-deceived. How, then, in general, can an agent deceive himself by successfully employing a self-deceptive strategy?\(^{20}\)

As an attempt to resolve these paradoxes, some philosophers adopt the Freudian

resolution which postulates separate mental subsystems, each of which functions independently. These philosophers suggest that self-deception is possible due to the fact that we are composed of subsystems or subagents, so that one mental subsystem employs the strategy on another mental subsystem without the latter knowing what the former is up to. Davidson, for instance, holds that there can be boundaries between parts of the mind. He believes that "the mind can be partitioned into quasi-independent structures that interact in ways that the Plato Principle cannot accept or explain."\textsuperscript{21} The Plato Principle states that no one intentionally acts against what one knows to be the best judgment. Even though the division of the mind is real and it is due to such a division that conflicting beliefs are possible, the boundaries between parts of the mind are not discovered by introspection. Unfortunately, Davidson does not have a detailed psychological account of quasi-independent mental structures.\textsuperscript{22}

David Pears offers a rather elaborate account of a functional theory of mental partitioning. Pears argues that within an agent there is a 'cautionary belief' that will help the agent avoid having an irrational belief, given his evidence. There is also a sub-system to which the 'cautionary belief' may be automatically assigned when it fails to intervene and stop the irrationality.\textsuperscript{23} He argues that the formation of motivated irrational beliefs that the agent is able to avoid forming can be explained psychologically by virtue of the intervention of the subsystem. Pears says that

The sub-system is built around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate centre of agency within the whole person, it is, from its own point of view, entirely rational. It wants the main system to

\textsuperscript{21}Davidson, 1982, pp.299-300.

\textsuperscript{22}Davidson, 1985, p.147.

\textsuperscript{23}Pears, 1984, pp.69-72, 84.
form the irrational belief and it is aware that it will not form it, if the cautionary belief is allowed to intervene. So with perfect rationality it stops its intervention.  

The mental partitioning approach faces many objections. The standard objection is, as Jeffrey Foss puts it, that it is not a resolution of self-deception but of the deception of one subsystem by another. Mele takes a different route. He argues that it is simply unnecessary to multiply mental systems in order to solve the problem of self-deception. Moreover, the deceivers need not be guided or monitored by the intention of deceiving themselves when they are engaged in the strategies of self-deception.

Mele puts a considerable amount of energy into criticising Pears' functional theory of mental partitioning. Mele points out that the hypothesis that an agent can deceive himself by letting one subagent employ a strategy while another subagent forms a belief may be true. However, Pears' account does not tell us how exactly psychological elements mediate between desires and avoidable irrational beliefs, when they are causally related. Second, even if it is true that the subsystem renders the cautionary belief ineffective, Pears should explain how it is done strategically. Third, there is a better alternative than that of Pears, namely, that we can identify the strategies of motivated doxastic irrationality without having to postulate Pearsian subsystems. For instance, empirical studies support the belief that an agent, motivated by his pertinent desire, may form an irrational belief and be self-deceived by virtue of biassed thinking and data manipulation.

Finally, Mele argues that the competence to avoid irrationality does not depend upon the

---

24 Pears, 1984, p.87.
25 Foss, 1980, p.239.
26 Mele, 1987, pp.139-144.
presence of cautionary belief. There are measures that can be taken to prevent self-deception, such as the honest questioning of one's own motivation in A-ing, the exercise of self-control, and so on. We may suppose that an agent who is in general capable of acting rationally but who A-ed irrationally during a particular period of time did so because, motivated by a certain desire, he failed to take preventative measures, rather than because of a cautionary belief being sent to the subsystem. Mele goes on to propose a different approach to solve the two paradoxes of self-deception. This approach involves two components.

First, Mele argues, we should not assume that self-deception needs to satisfy the doxastic condition which involves the agent's believing both that \( p \) and that \( \sim p \). Second, self-deception, although it involves intentional action, need not be intentional. That is, an agent does not have to have the intention to deceive himself in order to be self-deceived, even though successful self-deception requires that the agent intentionally acts in a certain way.

Mele argues that it is misguided to maintain that both deception and self-deception must be intentional. Mele insists that the notion of deception should not be conceived of as being always intentional. First, Mele argues that when we say that A deceived B, we say nothing about whether A is an intentional deceiver. The fact that we can be deceived by the appearance of the world suggests that a deceiver may have no intention in deceiving anyone. Second, self-deceivers typically do not deceive themselves with the intention of deceiving themselves. Third, although different ways of data-manipulation are strategies of self-deception, self-deceivers usually are engaged in these strategies without the intention of deceiving themselves. Self-deception is strategic in the sense that self-deceivers deceive themselves by behaving in

---

ways that involve data-manipulation. Finally, even if $A$ is a creature of intentionality, he may not know that $p$ is false at the time of deception. Without knowing that $p$ is false, $A$ may have intended to communicate $p$ to $B$ by telling him that $p$ and causing a belief that $p$ in $B$. And non-intentional deception is a case of deception nonetheless. In short, Mele maintains that it is not usually the case that the person enters self-deception with an intention to believe something he knows to be false. Sometimes the self-deceiver, motivated by his pertinent desire, only believes something against his best evidence. Self-deception is a display of desire-influenced irrationality, or, motivated irrationality. Mele's account of self-deception will be discussed in Chapter 2.

3. Desire and Motivation in the Theory of Self-Deception

Self-deception is both a state and a process. It is a state of motivated irrationality when one is self-deceived into believing that $p$, but one goes through a whole process of manipulating data relevant to the truth value of the belief that $p$ in order to get oneself deceived into believing that $p$. Such a process of data manipulation is not merely a disposition of the agent but is also composed of intentional actions. Therefore, the theory of action is relevant to self-deception in some important ways. In other words, the formation and retention of self-deception can be seen as a process of desire-action-belief. If self-deception is motivated irrational believing and intentional action is motivated action, then there is a parallel between a theory of self-deception and a theory of action in that both need to deal with the notion of motivation. For instance, a self-deceived person, motivated by desire, believes something false against his best evidence.

A weak-willed person, motivated by desire, intentionally acts against his best judgment.

Mele's account of motivation is a Humean account that treats every complete motivating state as a combination of belief and desire, with desire being essentially motivating states. I will deal with several issues concerning with desire and motivation in Chapter 3.

4. Related Phenomena

There are many phenomena related to self-deception. For instance, not only is self-deception similar to *akrasia* (i.e., acting against one's better judgment), but the two tend to reinforce each other. For instance, the cancer patient, due to her desire to be cancer free, explains away her symptoms in order to believe that she is cancer free. She does so against her better judgment. The success of explaining away her symptoms not only makes her a weak-willed person but also more convinced that she is cancer free.

Self-deception is also closely related to, if not a peculiar form of, self-induced deception. A spendthrift who tries to break his habit of over spending by intentionally manipulating his check book balance so that, at the end of each month, it will look to him as though he has less money than he actually has, may, as a result, spend less each month. If he perceives at some level that what shows in his check book is not true but still believes it, then he is, motivated by his desire to be economical, intentionally *deceiving* himself into believing something false. However, if he get his computer to automatically balance his money in such a way that it will look to him as though he has less money than he actually has the end of each month, and if he also forgets about his original plan, then we might say that he has induced

himself into believing something false rather than that he is self-deceived.

Self-deception is also related to wishful thinking in that they both involve motivational or evaluative elements, but they are different. The cancer patient is self-deceived in believing that she is cancer free by explaining away her symptoms. However, she is only a wishful thinker if she only tentatively explains away her symptoms and, once her explanations of her symptoms are challenged by further counter-evidence, readily accepts the belief that she has cancer.

Finally, self-deception is not as same as lying to oneself because the latter entails the intention to deceive oneself but the former does not. Again, the cancer patient is self-deceived in believing that she is cancer free by explaining away her symptoms, but if she is an experienced physician who clearly knows what her symptoms indicate and she still tells herself that she is cancer free, then she is lying to herself.

All these issues will be dealt with in Chapter 4.

I mentioned earlier that, even though self-deception may be viewed from different perspectives, for our present purpose I will treat it as a form of motivated irrationality. Irrationality can be understood by clarifying the concept of being rational. The concept of being rational differs among philosophers. Some philosophers, from the point of view of cognition, hold that to be rational one must conform to the laws of deductive logic. Others suggest that to be rational one must be able to judge based on appropriate inductive reasoning. Charles Daniels argues that self-deception is irrational because a self-deceived person believes that \( p \) without evidence that warrants this, because of a certain desire; or, because of a certain

---

\(^{30}\) Cohen, 1992, pp.415-419.
desire, he does not believe that $p$ in the face of evidence.\textsuperscript{31}

Some philosophers believe that to be irrational is to fail to give a reason for which one believes or acts (such as in the case of akrasia). Davidson holds that for a belief to sustain a contrary belief as in the case of self-deception, there must be a mental cause that is not a reason.\textsuperscript{32} Even though the process of self-deception can be explained by appealing to causally connected mental events, to cite such a causal relation of events and states is not to give the reason for which one acts or believes. It is by virtue of ignoring one’s own standard of rationality that self-deception is thought to be irrational.\textsuperscript{33} On the other hand, if the agent can, according to his own standard of rationality, be justified to mislead himself into believing something (even if the belief turns out to be false), then he is epistemically irresponsible without being irrational. I will discuss this issue in the last chapter as well.

Following a detailed exposition of Mele’s account of self-deception in Chapter 2, the chief goal of my thesis will be to show that Mele’s account is inadequate; in addition, I will show how his account can be improved by a reason explanation of self-deception that is based on a concept of deception narrower than the one that Mele has been using.

\textsuperscript{31}Daniels, 1974, pp.248-249.

\textsuperscript{32}Davidson, 1982, p.298.

\textsuperscript{33}Davidson, 1982, pp.297-298.
Chapter 2

Mele’s Theory of Self-Deception

Mele’s chief goal is to clarify the nature and etiology of self-deception. The following questions are what Mele focuses on in his attempt to explain self-deception:

1. Does a self-deceived person simultaneously possess the belief that \( p \) and the belief that \( \neg p \)?

2. Is garden variety self-deception intentional?

3. Should the study of self-deception be modelled on the study of other-deception?

4. Are Mele’s jointly sufficient conditions conceptually sufficient for self-deception?\(^1\)

5. What does it mean to say that self-deception is motivated self-misleading? For what reason does the self-deceiver manipulate data?

6. Is reason-explanation adequate for explaining motivated irrationality?

7. Why does the agent do \( A \) instead of doing \( B \) when he is free to choose to do either?

8. Is there a cure for motivated irrationality in general, or for self-deception in particular?

1. Two Paradoxes of Self-Deception

One of Mele’s chief tasks in his work on self-deception is to dissolve two paradoxes. By focussing on problems such as the clarification of the nature and etiology of self-deception and the process of self-deception, he tries to explain why and how self-deception happens. One

\(^1\)Mele, 1997b, p.127.
of the greatest difficulties in explaining self-deception is the paradoxical nature of the phenomenon. Basically we are faced with two paradoxes: the paradox of belief and the strategy paradox. The first paradox, the paradox of belief, is concerned with the following question: What does the self-deceived person believe, and what does he do given what he believes? The strategy paradox is concerned with the following question: How is self-deception possible? These two questions are interrelated in that we cannot satisfactorily answer one without answering the other.

The first paradox, the paradox of belief, is also the paradox of self-deception. It arises as follows:

For any $A$ and $B$, when $A$ deceives $B$ into believing that $p$, $A$ knows or truly believes that not-$p$ while causing $B$ to believe that $p$. So when $A$ deceives $A$ (i.e., himself) into believing that $p$, he knows or truly believes that not-$p$ while causing himself to believe that $p$. Thus, $A$ must simultaneously believe that not-$p$ and believe that $p$. But how is this possible?

One way of trying to solve this paradox is to argue that the nature of our beliefs is such that it is possible for one to hold a set of self-contradictory beliefs. The paradox of belief will no doubt dissolve right away if we can successfully demonstrate that one can simultaneously hold contradictory beliefs rationally. However, exploring the nature of belief is neither Mele's chief goal nor necessary for dissolving this paradox. Mele simply assumes that to believe that $p$ is to believe that the likelihood of $p$ is greater than 50%. He rejects the assumption that one must simultaneously hold a pair of contradictory beliefs in order to be self-deceived.

Some might argue that the behaviour of self-deceivers typically provides weighty evidence that they do simultaneously hold contradictory beliefs. If self-deceivers are not

---

2 Mele, 1987, p.121.

3 Mele, 1987, p.121.
characteristically in a peculiar doxastic condition (namely, believing both that \( p \) and that \( \neg p \)), how do we explain the behaviour of a person who sincerely avows that \( p \), yet systematically acts contrary to his sincere avowal? Mele responds to this challenge by maintaining that the conflicting behaviour itself does not provide strong evidence that one simultaneously holds contradictory beliefs. Instead of attributing inconsistent beliefs to the agent, in cases which the agent’s words systematically go against his deeds, we can say that he does not believe what he asserts. Mele argues that there is no ground to assume that an agent is in such a peculiar doxastic condition of simultaneously holding contradictory beliefs, when one is self-deceived. The paradox of belief is dissolved once we accept that the self-deceived person does not simultaneously hold the belief that \( p \) and the belief that \( \neg p \).

The second paradox of self-deception is the strategy paradox of self-deception. This paradox puzzles over the success of self-deceptive strategies employed by the self-deceiver. It goes like this. In general, \( A \) cannot successfully employ a deceptive strategy against \( B \) if \( B \) recognizes \( A \)'s intention and plan. In cases of self-deception, \( A \) and \( B \) are the same person. Since one always knows one’s own intention and plan, one cannot successfully employ a deceptive strategy against oneself. Therefore, self-deception should be impossible. However, people are in fact often self-deceived. How, then, in general, can an agent deceive himself by successfully employing a self-deceptive strategy?

Some philosophers have tried to solve this paradox by dividing the mind into different subsystems or subagents, each of which functions independently of the others. This approach

---


may seem to offer an easy way out of both paradoxes. Since the subsystems function independently of each other, one subsystem accepts the belief that \( p \), the other, that \( \neg p \). Moreover, one subsystem can successfully employ self-deceptive strategies without the other knowing it. However, this mental partitioning approach generates many further questions. For instance, who is the agent who is in charge of the subsystems, if there is one? How are different systems related to each other? Who mediates between the desires and beliefs? And so on. At the end of such probing, one might find that the mental partitioning approach has potential problems that are too troublesome to make it worth adopting.

Mele argues that for the purpose of explaining self-deception it is unnecessary to postulate subagents. First, Mele holds that the strategy paradox is generated by modelling self-deception on intentional other-deception, on the assumption that deception must be intentional. He rejects this assumption by insisting that the garden-variety cases of self-deception are not cases of \textit{intentional} deception, even if they involve intentional actions. Mele argues that when we say that \( A \) deceived \( B \), we say nothing about whether \( A \) is an intentional deceiver. The fact that we can be deceived by the appearance of the world suggests that a deceiver need have no intention to deceive anyone. Even if \( A \) is a creature of intentionality, he may not know that \( p \) is false at the time of deception. For instance, without knowing that \( p \) is false, \( A \) may have intended to communicate \( p \) to \( B \) by telling him something, which causes a belief that \( p \) in \( B \). Unintentional as it is, Mele says, \( A \) deceived \( B \) nevertheless.\textsuperscript{7} Second, Mele argues, self-deception seems to be motivated by desire and therefore it is \textit{non-accidental}, unlike non-intentional interpersonal deception, which may be accidental. Finally, instead of modelling

\textsuperscript{7}Mele, 1987, p.123.
self-deception on other-deception, Mele argues that we should pay attention to the parallel situation of non-intentionality between self-deception and akratic action: An agent who acts incontinently can act intentionally without intentionally acting incontinently; the self-deceiver gets himself into the state of self-deception with intentional actions, even though he does so without the intention to deceive himself.

As we can see from the above discussion, Mele has introduced two negative theses in order to dissolve the paradoxes: 1) the self-deceived agent does not simultaneously hold contradictory beliefs, and 2) the self-deceived agent need not deceive himself with the intention to deceive. The following are some of Mele’s positive claims:

First, being deceived entails having a pertinent false belief. This claim is a lexical one. Mele argues that when we say that a person is deceived about something, $p$, we often mean simply that he is in error with respect to $p$, or that he holds a false belief concerning $p$. If the belief turns out to be true, then he is not deceived in believing it.$^8$

Second, deception means misleading. Mele holds that $B$ is deceived by $A$ if $A$ misleads $B$ into believing something false. For instance, if $A$ misreads a word from a piece of written work to $B$, and $B$ believes something false as a result of such a mistake made by $A$, $B$ is said to be deceived by $A$, even though $A$ has neither motive for, nor intention, nor knowledge of making such a mistake or deception.$^9$ This claim indicates that Mele holds a rather broad concept of deception, one which may not be shared by others. For even though $A$’s misleading $B$ into believing something false is not justifiable, she is exculpable because she is not entirely responsible for her unintentional act; it is unfair to accuse $A$ of being a deceiver. I will further

---


discuss this issue in Chapter 3.

Third, in self-deception, the misleading must be motivated. Mele argues that, although it is *not usually* the case that the person enters self-deception with an intention to believe something that he knows to be false, unlike those cases of other-deception in which the deceiver need not have a motive, the self-deceiver’s believing something false is motivated by his pertinent desire. Mele argues that a *desiderative* element is necessary and unique to self-deception. That is, self-deception occurs because (in part) the agent *wants* something, and that if the agent wants to believe that $p$, it is because he wants it to be the case that $p$. Mele uses the terms ‘want’ and ‘desire’ interchangeably. In addition, he holds that to want something is to have some motivation to do (or to believe) something. Moreover, motivated believing in self-deception indicates that self-deception is non-accidental. Yet, non-accidental believing is not necessarily intentional believing. Without the intention to deceive or mislead himself, a self-deceiver can, motivated by a certain desire, intentionally do something in order to achieve that desire, and as result of such a motivated act, he believes something false.

Fourth, the self-deceiver usually believes that $p$ unwarrantedly. The self-deceiver usually believes a false belief against his best evidence. Modelling on his theory of *akrasia*, Mele holds that the self-deceiver believes incontinently because he believes intentionally, but that his believing incontinently is not part of his intention.

Finally, Mele holds that the relationship between the deceiver and the one who is deceived is one of cause. Self-deception is a display of desire-influenced, or motivated,

---

11 Mele, 1992, p.47.
irrationality. Therefore, self-deception can be understood as motivated by desire; one causes oneself to believe something false.

2. Self-Deception Described

Descriptions of self-deception can be divided into two kinds: strong and weak cases. Those who give the strong ones tend to believe that self-deception entails that the self-deceiver simultaneously believes both that \( p \) and that \( \neg p \). Some of them even hold that the self-deceiver must intentionally lie to himself. Mele holds that this kind of description of self-deception is inappropriate. It may either exclude too much or include too much. That is, it may either exclude cases which do not involve lying to oneself, or it may attribute such lying to every case of self-deception even though in some cases such lying is not obvious.

Mele adopts a weaker way of describing self-deception, under which the agent needs neither to be intentionally self-deceived, nor to be in a peculiar doxastic condition (namely, believing both that \( p \) and that \( \neg p \)). Garden-variety cases of self-deception, such as the cancer patient who ignores his symptoms or explains them away; the drunkard who will not admit his alcoholism in the teeth of the evidence; the mother who is blind to her son’s faults,\(^\text{13}\) and so on, can be explained by this weak description. Mele has proposed a characterization and jointly sufficient conditions for self-deception. He states that \( S \) is self-deceived with respect to the belief that \( p \) if

1) The belief that \( p \) which \( S \) acquires is false.

2) \( S \)'s desiring that \( p \) leads \( S \) to manipulate (i.e., to treat inappropriately) a datum or

\(^{13}\)Haight, 1980, p.82.
data relevant, or at least seemingly relevant, to the truth value of \( p \).

3) This manipulation is a cause of \( S \)'s acquiring the belief that \( p \).

4) If, in the causal chain between desire and manipulation or in that between manipulation and belief-acquisition, there are any accidental intermediaries (links), or intermediaries intentionally introduced by another agent, these intermediaries do not make \( S \) (significantly) less responsible for acquiring the belief that \( p \) than he would otherwise have been.\(^{14}\)

Condition 1) is lexical. By definition, to be deceived is to be in error; one is not deceived in believing that \( p \) if that \( p \) is true. Mele argues that there is a difference between entering self-deception in acquiring the belief that \( p \) (simpliciter), and entering self-deception as a result of acquiring the belief that \( p \). In the latter sense, the agent may enter self-deception by acquiring a false belief that \( r \) as a result of a mistaken inference based on a true belief that \( p \); or, he may enter self-deception by acquiring a true belief that \( p \) on the basis of unwarranted evidence, \( e \). Neither of these is a case of entering self-deception in acquiring the belief that \( p \) simpliciter. However, Mele argues that his phrase "in acquiring the belief that \( p \)" is to be understood as "in acquiring the belief that \( p \) simpliciter" hence he claims that one cannot be self-deceived into believing that \( p \) unless \( p \) is false.\(^{15}\)

Conditions 2) and 3) are concerned with the self-deceiver's contribution to his deception. They answer the question 'What must be added to false belief to yield a condition of self-deception?' Conditions 2) and 3) state that the agent's data manipulation is motivated by his desire relevant to the self-deceived belief, and that the relationship between his act (data


\(^{15}\)Mele, 1987, p.128.
manipulation) and one's belief acquisition is a causal one.

Finally, Condition 4) is present in order to limit the range to the pure and clear cases. It says that there should be no deviant causal chain which makes S believe that p. A deviant causal chain may look something like this: S desires p and believes that doing A will enable him to get p; but due to some external influence, S A-ed by accident rather than intentionally. In such a case, we say that S A-ed unintentionally because of the waywardness involved in the chain of desire-belief-action.

In his recent work, “Real Self-deception,” Mele has modified his characterization and his set of jointly sufficient conditions of self-deception. They now look like this:

1) The belief that p which S acquires is false;
2) S treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way;
3) This biased treatment is a nondeviant cause of S's acquiring the belief that p;
4) The body of data possessed by S at the time provides greater warrant for \( \neg p \) than for p.

The differences between the old and the new sets of sufficient conditions are the following: In the case of Condition 2), the difference between the two sets is merely verbal. For Mele, to be caused by one’s desire to A is to be motivated to A. Condition 3) in the new set is a combination of conditions 3) and 4) in the original set. Only condition 4) in the new set is an additional condition which was not included in the old set. However, Mele holds that condition 4) in the new set is not a necessary condition of self-deception. The reason is that S's ignorance of the data warranting \( \neg p \) may or may not be the result of his motivationally biased treatment. S may be ignorant of the data supporting \( \neg p \) because he was motivated to remain so
ignorant. If this is the case, S is self-deceived in believing that $p$. On the other hand, S may be deceived in believing that $p$ because he was ignorant. But this is not a case of self-deception. Most of my work has been based on Mele’s original characterization and jointly sufficient conditions of self-deception. Since his new set is basically consistent with his old set, the rest of my discussion will still focus on the old set.

The chief virtue of Mele’s characterization of self-deception is that it allows us to explain the phenomenon without the supposition that the agent simultaneously believes that $p$ and that $\neg p$.\textsuperscript{16} Since one can be self-deceived into believing that $p$ without believing that $p$ is false, what happens in these cases is that, instead of simultaneously believing that $p$ and that $\neg p$, S deceives himself by preventing himself from holding a certain true belief. For instance, motivated by a certain want S has, S believes that $p$ by shifting his attention away from data ($d$) which can be counted against the belief that $p$. There is no reason for us to think that S must do so knowingly.

Mele maintains that even though the person’s entering self-deception does not involve a peculiar doxastic condition of simultaneously believing both that $p$ and that $\neg p$, self-deception is nevertheless a troublesome phenomenon. The real problems with self-deception are parallel to those of akratic action, since they both are desire-influenced irrationality.

3. Entering Self-Deception

a. Entering Self-Deception by Acquiring a Belief

Mele points out that the paradoxes of self-deception are generated on the entering into

\textsuperscript{16}Mele, 1987, p.128.
of a condition of self-deception, as opposed to while being or remaining in self-deception.\textsuperscript{17} If a false belief that \( p \) is acquired by means of data-manipulation, the self-deceiver’s doxastic condition must be quite different from that of non-self-deceivers. We do not have to assume that it must be a peculiar doxastic condition of simultaneously believing both that \( p \) and that \( \neg p \). By examining the entering condition of garden variety cases of self-deception, Mele tries to answer the question ‘How is self-deception possible?’ without committing himself to a peculiar doxastic condition of simultaneously believing both that \( p \) and that \( \neg p \).

According to Mele, self-deception is a matter of desire-influenced, or motivated, irrationality. He suggests that there are four common ways (or strategies) that an agent may be motivated to believe something false. Mele sometimes calls them the ‘hot’ mechanism:

1) Negative Misinterpretation: the agent, led by his desire that \( p \), misinterprets as not counter evidence data that he would easily recognize as evidence counting against \( p \) in the absence of the pertinent desire.

2) Positive Misinterpretation: the agent, led by his desire that \( p \), misinterprets as supporting evidence data that he would easily recognize as evidence counting against \( p \) in the absence of the pertinent desire.

3) Selective Focussing/Attending: the agent, led by his desire that \( p \), fails to focus his attention on evidence that counts against \( p \) and focuses instead on evidence suggestive of \( p \). He would easily avoid such selective focussing in the absence of the pertinent desire.

4) Selective Evidence-Gathering: the agent, led by his desire that \( p \), both overlooks easily obtained evidence for \( \neg p \) and finds evidence for \( p \) which is much less accessible. He

\textsuperscript{17}Mele, 1987, p.125.
would easily avoid such selective evidence-gathering in the absence of the pertinent desire.\textsuperscript{18}

One might argue that even if in most cases of self-deception the agent does not simultaneously hold a pair of contradictory beliefs, he must at least believe that his data do hold against \( p \) since he misinterprets data. If so, then the paradox of belief still holds. Mele thinks that this does not need to be the case. It may just be that the agent recognizes that data might be held against \( p \). Other than misinterpretation of the data, given his emotional condition, he might intentionally or unintentionally shift his attention away from the data, or, he might see the data in a certain way which leads to his belief that \( p \).

Mele argues that even intentional misinterpretation does not necessarily involve the peculiar doxastic condition of simultaneously believing both that \( p \) and that \( \neg p \), although misinterpretation need not be intentional. In cases of intentional misinterpretation, the agent believes that the data supports \( \neg p \), but he attempts to construct the data in a certain way so that they look like evidence for \( p \). By the time he forms the belief that \( p \), he has already abandoned the original belief that \( \neg p \).

Mele also points out a difference between unintentional positive misinterpretation and intentionally putting a charitable construction upon the data. Both unintentional positive misinterpretation and the intentional putting of a charitable construction upon the data can lead to self-deception. But in the latter case, the agent does not have the intention to misinterpret the data. The difference between the two may be shown in the following two examples that Mele gives.

Example 1.

\textsuperscript{18}Mele, 1987, pp.125-126.
A woman who intentionally takes an 'upbeat' approach in interpreting the data concerning her husband's fidelity may want to exercise the virtue of charity and believe that she may do so in this way.\textsuperscript{19}

Example 1. is one of intentional charitable construction of data. The next example is one of unintentional positive misinterpretation:

Example 2.

Suppose, ..., Sid is very fond of Roz, a young woman with whom he often eats lunch. If he wants it to be the case that Roz loves him, he may interpret her refusing to go out on dates with him and her reminding him that she is very much in love with her steady boyfriend, Tim, as an effort on her part to 'play hard to get' in order to encourage Sid to continue to pursue her.\textsuperscript{20}

The first example shows that intentional charitable interpretation may lead one to intentional self-deception without the intention of misinterpretation; the agent's intention is to be charitable. The second example shows that misinterpretation occurs when one is over-positive in one's interpreting certain data without realizing or even intending it. In either case, Mele points out, there is no reason to believe that the agent must be in the peculiar doxastic condition when the agent enters the self-deception. Neither unintentional positive misinterpretation nor intentionally putting a charitable construction upon the data need involve simultaneously holding contradictory beliefs.

Although garden-variety cases of self-deception are not intentional, intentional self-deception does occur. Even so, intentional misinterpretation does not entail that the agent simultaneously holds contradictory beliefs. Mele argues that it takes time, no matter how brief, for one to construct the data differently for a believer to get himself to give up the belief that $\neg p$

\textsuperscript{19}Mele, 1987, p.130.

\textsuperscript{20}Mele, 1987, p.126.
(which is warranted by the data) and to believe that \( p \) (which is supported by misinterpretation of the data).\(^{21}\) Here is an example of intentional self-deception that does not put the agent in a peculiar doxastic condition. \( S \) makes a plan to deceive himself about a belief that \( p \). He believes that \( \neg p \) on the basis of \( e \). However he wants \( q \) and he believes that if \( p \), then \( q \). He decides to deceive himself into believing that \( p \) so that he may have \( q \). When he succeeds in believing that \( p \), he stops describing the plan as a self-deceptive one. His believing that \( \neg p \) precedes his believing that \( p \); therefore, he is not in a peculiar doxastic condition at the time he enters self-deception.

\[ b. \text{Entering Self-Deception by Retaining a Belief} \]

Mele argues that the process of entering self-deception by retaining a belief can be understood in a very similar way to that of entering self-deception in acquiring a belief. Both processes have to do with a desire-influenced inappropriate treatment of data by means of negative or positive misinterpretation, selective focusing and selective evidence-gathering. The difference in the case of the process of entering self-deception in retaining a belief is that the agent keeps a belief that is no longer true instead of acquiring a false belief. The belief was true in the past, but now it is false because the facts have changed. However, the agent's desire to believe it remains unchanged and he may exploit the data-manipulating mechanism (described in the previous passages) in order to retain this belief. Because the agent's believing it is motivated by pertinent desire, because his believing it is arrived at by the exercise of data-manipulation, and because the retained belief is now false, the agent has entered into self-

\(^{21}\)Mele, 1987, p.129.
deception by retaining a belief.\textsuperscript{22}

4. The Strategies of Self-Deception

Let us review the strategy paradox of self-deception once more. In general, \( A \) cannot successfully employ a deceptive strategy against \( B \) if \( B \) knows \( A \)'s intention and plan. In cases of self-deception, \( A \) and \( B \) are the same person. Thus self-deception seems impossible because one always knows one's own intention and plan. However, in fact people are often self-deceived. How in general can an agent deceive himself by successfully employing a self-deceptive strategy?

Mele's answer is this. The four ways of motivating one to form a false belief (i.e., the 'hot' mechanism) are self-deceptive strategies that an agent can successfully employ. They can be gathered into two groups: the internal-biassing strategies and the input-control strategies.

1) Internal-biassing strategies are constituted by the manipulation of data that one already has. They include A) positive and negative misinterpretation of data that one already has, and B) selective focussing or attending to data that one already has.

2) Input-control strategies consist of the agent's controlling which data he acquires. They are A) strategies that operate upon data already present in the world (such as, selective data-gathering), and B) strategies that control the input of data by generating data (such as, "acting as if").

Moreover, there are three different ways of employing a self-deceptive strategy:

1) the non-intentional employment of the strategy;

\textsuperscript{22}Mele, 1987, p.131.
2) the intentional employment of the strategy without the intention of deceiving oneself;
3) the intentional employment of the strategy with the intention of deceiving oneself.

a. Internal Biassing Strategies

Mele quotes some of the common sources of unmotivated irrational beliefs that have been identified by the social psychologists Nisbett and Ross. To explain how one is motivated into self-deceived believing that \( p \), Mele holds that the internal-biassing strategies and the input-control strategies exploit the common sources of unmotivated irrational beliefs, and that such exploitation is further promoted by one’s desire.

First, let us look at the common sources of unmotivated irrational beliefs, Mele sometimes calls them the ‘cold’ mechanism:

1) Vividness of Information: The subjective vividness of a datum is often a function of the subject’s interests. Vivid data tend to influence one’s belief formation and belief retention disproportionately.

2) The Availability Heuristic: People’s estimation and assessment of an event are often influenced by the accessibility of the event in their perceptual and thinking processes. When perspectives are altered, the same viewer can correspondingly alter his assessment of a single event.

3) The Confirmation Bias: People tend to search, when testing a hypothesis, for confirming instances more often and more readily than for disconfirming ones.

4) Tendency to Search for Causal Explanations: People tend to search for causal

\[\text{For more details, see Nisbett and Lee, 1980.}\]
explanations of events. Given 1), 2) and 3), the causal explanation we give for ordinary events in our daily life may be illfounded, and we may be endorsing more hypotheses than we ought.\textsuperscript{24}

Mele states in Condition 3) of his jointly sufficient conditions of self-deception that $S$’s manipulating data relevant to the truth value of $p$ is a cause of $S$’s acquiring the belief that $p$.

This condition raises the following question: How, in general, can $S$ be self-deceived by manipulating data? This question can be broken down to three further questions: 1) How can one’s employing a deceptive strategy $X$ result in his believing that $p$? 2) How can one’s desiring that $p$ motivate him to employ a deceptive strategy $X$? 3) How can one’s employment of a deceptive strategy $X$, motivated in the manner that it is, \textit{fail} to be self-defeating? Each of these three questions will be answered in turn.

First, how can one’s employing a deceptive strategy $X$ result in his believing that $p$?

Mele argues that internal biasing strategies exploit the ‘cold’ mechanism (the common sources of unmotivated irrational beliefs) without the agent’s intention to render them effective. For example, when the agent manipulates data by exercising an internal biasing strategy such as positive misinterpretation, his misinterpretation can have a strong impact on the availability and vividness of data. His positive misinterpretation may also suggest a favoured hypothesis, thereby promoting a biased search for confirmatory data and further misinterpretation.

Second, how can one’s desiring that $p$ motivate him to employ a deceptive strategy $X$?

Mele argues that one’s desire may motivate one to employ a deceptive strategy in a variety of ways. For instance, the data supporting $p$ may appear to be more vivid to the agent because of his desire that $p$, so that the supporting data dominate his attentional field. Or, the agent’s

\textsuperscript{24}Mele, 1987, pp.144-145.
desire that \( p \) may affect the hedonic quality of thoughts about different data, so that he decides to focus on confirmatory data rather than disconfirmatory ones.

Finally, how can one’s employment of a deceptive strategy \( X \), motivated in the manner that it is, fail to be self-defeating? This is a problem only in cases of intentional self-deception. It is only when the agent manipulates data with the intention of deceiving himself that the formation of his deceptive belief should be undermined by his realization that he has manipulated his data. One way of explaining this is that intentional self-deception takes time to achieve. By the time one successfully employs the self-deceptive strategies and gets oneself to believe in certain ways, one’s perspective has already changed so much that one might not view one’s past intention to deceive oneself as such.\(^{25}\)

It is worth noting that Mele makes the causal relation between one’s desire and one’s employment of self-deceptive strategies appear to be rather weak. The second answer above says that one’s desire may motivate one to employ a deceptive strategy in a variety of ways. The third answer above says that self-deceptive strategies are not monitored or guided by the intention to deceive oneself. I will argue that the relationship between the agent’s desire and his employment of deceptive strategies may be tighter than what Mele has made it out to be. There is no reason to assume that these strategies may not be guided by the agent’s intention to conceal or falsify the truth value of the opposite belief (that \( \sim p \)). And if the intention to conceal or to falsify comes in shortly before \( S \)’s data manipulation, then this intention is not only antecedent to his data manipulation, but also an effective cause of it. I will discuss this issue in the next chapter.

\(^{25}\)Mele, 1987, p.149.
b. Input-Control Strategies

(1) Input-Control Strategies Operating on Readily Available Data

Input-control strategies consist in the agent's controlling which data he acquires. These strategies can be further divided into two types: those that operate on readily available data and those that generate data. The strategies that operate on readily available data are similar to the internal-biassing strategies that we have discussed before. That is, they are all constituted by the manipulation of data that one already has. They can employ strategies such as positive and negative misinterpretation of data that one already has, or selective focussing or attending to data that one already has. Moreover, the answers to the questions of efficacy and motivation of strategies self-deception are similar to those as well. This will be made clearer by the following three points.

First, let us look at the efficacy of S's employment of input-control strategies operating on readily available data. How can S's employment input-control strategies operating on readily available data result in his believing that \( p \)? Mele believes that selective focussing and evidence-gathering (as part of the input-control strategies operating on readily available data) have significant impact on the availability of data (i.e., their accessibility in one's perception or memory), which in turn has a great influence on belief-formation.

Second, how can one's desiring that \( p \) motivate him to employ input-control strategies operating on readily available data? The answer is that non-intentional input-control strategies operate on readily available data because the vividness of certain data is supportive of \( p \) due to the agent's desire that \( p \). In other words, a desire causes certain data to appear more salient to the agent. It is also possible that one's desire-motivated choice of companions may result in
one's selective exposure to data supportive of the hypothesis one desires. For instance, S feels more comfortable being with people who believe that p because S wants p to be true. In addition, S's motivated choice of friends may lead him to select more data supportive of the belief that p. In the cases of the intentional employment of these strategies, S's desire affects his hedonic quality of attention to memories, which results in his deciding to focus only on data which is more pleasant for him. For instance, if S wants p he may find evidence e unpleasant to attend to because e appears to be unsupportive of the belief that p. Consequently, S intentionally focuses his attention on data contrary to e.

Finally, how can one's employment of input-control strategies operating on readily available data motivated in such a manner, fail to be self-defeating? Mele insists that, in most cases, the agent non-intentionally deceives himself by intentionally or non-intentionally engaging in these strategies. Engaging in self-deceptive strategies, intentionally or not, does not entail knowing one's deceptive plan. Therefore, unless the agent manipulates the input-data with the intention of deceiving himself, the problem of self-defeating (i.e., the problem that the agent's knowledge of his own plan should undermine the success of his self-deceptive plan) does not apply. However, when intentional self-deception happens, the agent's recognition of his deceptive plan would undermine his belief-generating process by weakening his biased data-base. But there is still no need to assume that only mental partitioning can solve this problem. A better solution lies in the idea that self-deception takes time to accomplish and in the recognition of the fluidity of one's memory. The agent may interpret his plan differently at different stages of his self-deceptive project, and in so doing, his employment of input-control

strategies operating on readily available data fails to undermine his project.

(2) Data-Generating Input-Control Strategies

Now let us answer the three questions with respect to the efficacy and motivation of employing data-generating input-control strategies. First, how can one’s employing data-generating input-control strategies result in his believing that \( p \)? One of the data-generating input-control strategies that Mele introduces is called ‘acting as if.’ The agent’s acting as if \( p \) can generate data supportive of \( p \) by both intrapersonal and interpersonal behavioural observation. \( S \) acts as if \( X \) and his behaviour of \( X \)-ing will be observed by himself, thereby being treated as part of his evidence for his \( X \)-ing or being \( X \). \( S \)’s acting as if \( X \) may also make his friends perceive and treat him in a way that will make him more certain about his being \( X \).

The second question with respect to the motivation of employing data-generating input-control strategies is, how can one’s desiring that \( p \) motivate him to employ data-generating input-control strategies? The above discussion indicates that acting as if may lead to self-deception without \( S \)’s intentionally acting as if. Even when \( S \) intentionally acts as if \( X \), he is not intentionally deceiving himself because he can be ignorant of what is motivating him to act as if \( X \). We can say that \( S \)’s employment of the strategy of acting as if is motivated if it is intentional. That \( S \)’s desiring \( X \) motivates him to act as if \( X \), and that our tendency to form beliefs about ourselves on the basis of the observation of our behaviour, make acting as if a good strategy for arriving at self-deception.\(^{27}\)

Finally, how can one’s employment of data-generating input-control strategies,

motivated in the manner that it is, fail to be self-defeating? Mele offers the following three examples to show that $S$ intentionally acts as if without such strategies being undermined by his knowledge of such an act. Example (1): $S$ wants to be $X$, and $S$ believes that acting as if he were $X$ may lead him to achieve $X$. Given this belief and his desire, $S$ has a reason to act as if he were $X$. Example (2): $S$ values $Y$ but he believes that he does not have $Y$. He enjoys his actions that are associated with $Y$. Such actions indicate to him that he is making progress toward having $Y$ and consequently he is motivated to act as if he had $Y$. Example (3): $S$ is unhappy with $T$ but $S$ still has to deal with $T$. $S$ may act as if he is content in order to make his dealing with $T$ easier.\(^{28}\)

5. Reasons and Motivation

a. Reason-Explanation

Thus far, we have discussed Mele's causal account of self-deception. He has given us some ideas of how (entering and retaining) self-deception takes place, by appeal to motivated and unmotivated self-deceptive mechanisms. Mele's explanation of the agent's deceptive bit of data-manipulation tells only half of the story. The other half of the story has to do with the reason-explanation of self-deception. What are the reasons for the agent to be self-deceived? Davidson points out that the process of self-deception can be explained by causally connected mental events. But such a causal chain cannot be justified by the agent's own standard of rationality, and this is why self-deception is thought to be a form of irrationality.\(^{29}\) However, I would argue, it is not so settled that the self-deceived person does not have a reason to deceive

\(^{28}\)Mele, 1987, pp.157-158.

\(^{29}\)Davidson, 1982, p.298.
himself. Hence it is too early to say that self-deception is necessarily irrational.

Mele is not against giving a reason-explanation for self-deception. However, he points out that, traditionally, it is believed that a want/belief pair constitutes a reason for action only when the belief-element is conducive to achieving what is wanted. This traditional belief-desire model of action-explanation works fine for dealing with rational action, but it becomes inadequate in the case of explaining an agent’s A-ing when the agent has competing reasons for an action, such as in cases of akrasia, or akratic believing. For example, S A-ed incontinently. That is, S judged that it was better to B but he went ahead to A. We want to know why S A-ed rather than B-ed. What is the reason for S’s A-ing rather than B-ing? There is no reason, Mele says. Mele promises to offer an account that will enable us both to give reason-explanations and to supplement what reasons-explanations are unable to explain. The supplementary explanatory items include positive/negative/total motivational bases of desires, the agent’s attentional condition, the perceived proximity of a reward, self-control, etc. Mele believes that it is only when we look at the explanatory items other than reasons that we start to understand why an agent acted for the reasons that he did rather than for his competing reasons.30

Similarly, Mele argues, when it comes to self-deception, we find that the self-deceiver may believe that data q counts against p. But this belief is not one “about the conduciveness of some behaviour to the achievement of the object of a desire.”31 In the case of intentional self-deception, S intentionally As (manipulates the relevant data), not for the reason that the data counts against p, but instead, for the reason that S wants to achieve p and believes that he may

achieve $p$ by $A$-ing.$^{32}$

Finally, Mele argues that self-deception is irrational because the self-deceiver, due to his desire, believes against his better evidence, or against the better evidence he would have had, or could easily have acquired, if he did not have the desire in question.$^{33}$ I will argue in the next chapter that this kind of reason-explanation is not the right one for dealing with self-deception. I will also argue that there are appropriate and inappropriate manipulations of data, so that data manipulation is not intrinsically deceptive, even if it is motivated. I will argue that only when one attempts to conceal or falsify the truth value relevant to the belief that $p$, can we say that one has committed deception. Does this mean that the paradox of deception still holds? Not necessarily so. Only when $S$ manipulates data for the reason of concealing or falsifying the truth relevant to the belief that $p$ can we say that $S$ is deceiving himself. Such data-manipulation takes time. As long as there is a time-gap, there is no paradox. Finally, although self-deception is epistemically irresponsible, it is not necessarily irrational.

**b. Motivation and Motivational Strength**

In this section, I will discuss Mele’s account of motivation of action. Although action is very different from belief, there is a parallel between the two. It is easier for us to understand the nature of motivated irrational believing once we understand the nature of motivated irrational action.

How do we understand the claim that $S$'s data-manipulation of the truth value of $p$ is motivated by his desiring that $p$? Specifically, how shall we conceive the nature of motivation?

$^{32}$Mele, 1987, p.130.

According to Mele, motivation is the action-initiating mental state. Its conative force is essential for intentional action to the extent that intentional action is motivated action. However, it is possible for someone to act intentionally without being motivated to so act. One can be extrinsically motivated to act intentionally. For instance, \( S A \)-ed intentionally because he was (intrinsically) motivated to \( B \) and believed that his \( A \)-ing was conducive to \( B \).

Mele maintains that the relationship between motivation and intentional action is not that reason is a slave of the passions, but rather, that reason has a motivational reservoir of its own. So, 'intentional action is motivated action' should be interpreted in this way: Everything that we intentionally do we are motivated to do, but what motivates \( A \)-ing is not necessarily one's motivation to \( A \) but one's motivation to do something which one believes will result from one's \( A \)-ing.\(^{34}\)

The next question that comes to us is, when \( S \) has competing motivations to do \( A \) or \( B \) while \( A \) and \( B \) are incompatible, why did \( S A \) rather than \( B \) when he was motivated to both incompatible actions \( A \) and \( B \)? Here, the notion of motivation alone is insufficient to explain an agent's action when he has conflicting motivations. One answer is that the agent chose to do \( A \) rather than \( B \) because he had judged it better to do \( A \), but this answer is challenged by the existence of akratic action, in which the agent judges it best to do \( A \) yet does \( B \) intentionally and without compulsion. Another answer to the above question is that the agent \( A \)-ed because he was most motivated to \( A \) at the time of his \( A \)-ing. However, the notion of motivational strength has been questioned by some philosophers. For instance, Irving Thalberg argues that to say that \( S \) is most strongly motivated to \( A \) at \( t \) is simply to say that he does \( A \) at \( t \). Therefore, the

notion of motivational strength has no explanatory power.  

Mele supports the idea that motivation to act is always a matter of degree. He argues that the notion of motivational strength is a useful one in that it provides "a point of departure for one who is seeking an explanation of S's A-ing". By explaining the agent's motivational condition at the time of his action, we can explain his acting. Here, Thalberg's answer is not very helpful. If the sole criterion of relative motivational strength is what the agent does, then the motivational strength of his wants cannot explain what he does without avoiding circularity. Mele argues that there are indicators of motivational strength other than the agent's subsequent action, such as the agent's own reports of his past experience of making similar choices which were effective under similar circumstances. Mele says,

Indeed, an agent's present behaviour concerning temporally distant action-goals is generally a better indication of the relative strength of his present motivation to perform the future actions than is his subsequent performance of one of the competing future actions.  

However, we are not concerned that the agent was most motivated to A, but rather, why he was most motivated to A rather than to the competing action, B. At this point, Mele argues, the agent's balance of motivation at the time of action has to be taken into account. Mele holds that there is want in the evaluative sense and want in the motivational sense. What the agent wants more in the evaluative sense may not be what he wants more in the motivational sense, and vice versa. When S has conflicting wants A and B, he may judge that it is better to A and therefore forms an intention to A on the basis of evaluation. That he ranks A over B does not mean he is more motivated to A because "any ratiocinative weighing of wants typically is not a

---

35Thalberg, 1985, pp.88-103.

weighting in respect of motivational force, but rather a ranking in respect of value."

Mele further argues that there is an issue of whether one’s motivational force for an action is in alignment with one’s ranking of it. Suppose that S intended to either A or B but not to do both at the same time. If S believed that, all things considered, it was better to A than to B, and if S’s motivational force for A and B was in alignment with his evaluation of them, then S would have A-ed rather than B-ed. On the other hand, if S’s intention to A was formed on the basis that all things considered it is better to A than to B, and if the motivational force for A and B at the time was not in alignment with his ranking of them, then he would have B-ed rather than A-ed.

Finally, Mele argues that in a case of akratic action, S has reasons for doing A and B, and all things considered, S judges that A is the best for him to do. But he intentionally does B. Why does S B instead of A? To cite the reasons for B-ing will not answer the question since there are reasons for A-ing as well. So the real question is, why does S act on this reason rather than the other? Here, Mele says, is where the reason-explanation comes to an end because S A-ed for no reason. Factors other than reasons that can explain why S A-ed have to be brought into account. One such factor is self-control.

Before we go on to discuss self-control, let us consider the relationship between the above discussion of the balance of motivation and how it can be used to explain self-deception. Mele’s theory of action is applicable to akrasia as well as to self-deception. First, we can explain why S is self-deceived by citing his motivation for the deception, and we can also explain why S was self-deceived in believing that p when he also had reason for believing that


\(~p\) by citing factors other than reasons. Second, most cases of self-deception are cases of unwarranted believing. That is, motivated by certain desire, S manipulates data relevant to the truth value of the belief that \(p\) in order to get himself to believe that \(p\), regardless that good evidence for believing \(~p\) is available to him. The parallel between akratic action and akratic believing is striking. For one thing, the irrationality of both akrasia and self-deception is causally influenced by our desires, and such causal influence by our desires may not be entirely under our control, even though exactly how the causal connection works is not clear.\(^{39}\)

6. Self-Control, Continent Action and Balanced Motivation

Mele argues that the agent who employs self-deceptive strategies, without the intention of deceiving himself, acts freely. Even though the readily available evidence supports the self-deceiver’s favoured hypothesis when he employs self-deceptive strategies, he would not be self-deceived had he taken some preventive measures. Now we are about to discuss Mele’s solution to cure our motivated irrationality.

Mele suggests certain remedies for motivated irrationality, namely, the self-control of the autonomous agent and ways to achieve it. Self-control is, according to Mele,

the ability to master motivation that is \textit{contrary} to one’s better judgment -- the ability to prevent such motivation from resulting in behaviour that is contrary to one’s decisive better judgment\(^{40}\)

This ability includes both strength and skill of resistance to motivation that is contrary to one’s decisive better judgment. A self-controlled person must have learned self-control techniques and be able to execute them skilfully. That means, someone who merely acts according to his


\(^{40}\)Mele, 1987, p.54.
decisive better judgment without knowing any self-controlled technique, or who has learned some self-control techniques but does not know when it is appropriate to employ them, is not a self-controlled person.

Mele argues that, even though our beliefs are not under direct voluntary control as our actions are, they are nevertheless under our indirect control. Our decisions on how to gather data, for instance, can influence what we come to believe. But how is doxastic self-control possible if the control is only indirect? This is certainly a more urgent issue for our purpose since we are dealing with self-deception, which is more of a doxastic problem than a problem of action. The first answer is that a self-deceived person is typically believing against his better evidence. It is typically a form of akratic believing. A self-deceived person has failed to make a better judgment -- a judgment that is better without qualification. It is a better judgment in the sense that it overrides any competing judgment that one may hold. Mele uses the term 'self-commitment', “a commitment of the self by the self (rather than by some principle that is not the self’s),” to denote such evaluative overridingness. Mele holds that doxastic self-control is “the ability to master motivation that is contrary to a doxastic self-commitment (and to the judgments that generate such a commitment), whether the commitment be to a way of assessing, gathering, or attending to evidence, or of holding (or not holding) a certain belief”. That is, one cannot have doxastic self-control unless one takes doxastic better judgment to be overriding any other competing judgment. A doxastic self-controlled person must have learned doxastic self-control techniques and is able to execute them skilfully. He is appropriately

---


motivated to believe as he judges best in the doxastic realm and has the strength and skill to resist the motivation to believe the contrary.\footnote{Mele, 1987, p.60.}

According to Mele, continent behaviour varies in degrees. Arguably, the highest grade of continent believing is to believe on one’s better epistemic judgment due to one’s exercising self-control. The methods of such control that Mele has suggested are basically ways of behavioural modification that are widely practised in cognitive psychotherapy.\footnote{This gets into the area of cognitive psychotherapy which is far beyond the scope of my present work. For further reading of the topic, Mele recommends Brandt, 1979.} Mele claims that the highest grade of continent behaviour is that which acts on the agent’s better judgment that is \textit{due to} his exercising self-control. The next grade of continent behaviour is that which acts \textit{on} the agent’s better judgment -- intentional behaviour guided by the agent’s own judgment without the exercise of self-control. The lowest is that which acts merely \textit{in accordance with} one’s better judgment -- intentional behaviour guided by rules that the agent recognizes. Mele maintains that once we have examined continent actions, we can easily see how the balance between judgment and desire works.\footnote{Mele, 1987, pp.99-100.}

\section*{7. Conclusion}

Mele has addressed all the issues that I raised at the beginning of this chapter. Mele’s chief goal is to understand the nature and etiology of self-deception. He claims that self-deceived persons are ordinary rational people and that self-deception is neither paradoxical nor mysterious. Self-deception can be explained without the help of exotic concepts such as mental
partitioning or unconscious beliefs. Mele argues that the paradoxes of self-deception are generated by false assumptions, such as that the self-deceived person must simultaneously believe that $p$ and that $\neg p$ and that he must deceive himself intentionally. Having these false assumptions is the result of the misguided approach of modelling self-deception on other-deception. Mele claims that self-deception is more similar to *akrasia* than to other-deception, in that both self-deception and *akrasia* are forms of motivated irrationality. Mele claims that he has dissolved the paradoxes of self-deception, especially the strategy paradox, by giving a psychological account of the process of self-deception. In the end, Mele has also offered self-control as one of the remedies for self-deception.
CHAPTER 3

A Critical Examination of Mele’s Theory

Mele has made a number of contributions to the philosophical study of self-deception. First, he has dissolved two paradoxes of self-deception. Second, he has rightly argued that self-deception does not require the intention to deceive oneself. Third, he has made good use of empirical studies in psychology to explain the biased thinking which may lead to self-deception.

On the other hand, there are some problems with his account. Mele’s notion of deception is problematic. Mele’s account of self-deception is neither sufficient nor inclusive enough. Mele uses desire as the only motivational force in formation and retention of self-deception, and this, I will argue in a later part of this chapter, makes Mele’s naturalistic account of self-deception philosophically unsatisfying.

1. Deception and Self-Deception

I take it that self-deception is a form of deception and that the similarities between self-deception and other-deception are quite striking. For instance, both self-deceivers and other-deceivers are motivated misleaders. Both self-deception and other-deception are grounded in communication. The communicational ground in self-deception is one’s ability to rationalize, or, to adopt van Fraassen’s term, one’s ability to think sotto voce.¹ Moreover, both self-deception and other-deception involve some kind of error. It is hard to deny that ‘to be deceived’ can mean ‘is in error’, but the error may be found in the belief held by the agent, it

¹ van Fraassen, 1988, p.125.
may be in the process of the agent’s acquiring a particular belief, or, it may be in both. A more interesting question is what kind of error plays a central role in self-deception. In addition to the above similarities between the two types of deception, there are other common features shared by both and including, but not limited to, the following:

1) There is something sneaky (which requires some special cognitive efforts in order to hide something) on the part of both kinds of deceiver;

2) Deceivers of both kinds believe, or at least are aware that, the opposite belief (i.e., the belief that \( \sim p \)) is, or may be, the case;

3) Deceivers of both kinds have a goal, and believe that getting the victim to believe that \( p \) is conducive to that goal;

4) Deceits of both kinds involve concealment or falsification of what the deceiver takes to be relevant to the truth of that \( p \);

5) This deceptive behaviour is motivated by a) having a goal, b) believing that getting the victim to believe that \( p \) is conducive to that goal, and c) believing or at least being aware that, the opposite belief (the belief that \( \sim p \)) is or may be the case;

6) Both self-deception and other-deception attribution face the same question, namely, should the attribution of deception depend more on one’s deceptive intention and deeds than on the truth value of the relevant belief, or the other way around?

The following questions will be dealt with in searching for an appropriate notion of deception:

1) What is deception?

2) What distinguishes deception from other errors?

3) What is the ground for deception (what makes deception possible)?
4) Does deception necessarily involve a false belief?

I will deal with each question in due order in the following sections.

1.1 To define deception is by no means simple. Mele’s notion of deception is that to be deceived is to be in error. He says:

When we say that a person is deceived about something, X, we often mean simply that he is in error with respect to X, or that he holds a false belief concerning X. And there is a corresponding natural use of the verb ‘deceive’ in which to say that A deceived B with respect to p – something which is typically done by inducing a false belief that p in B. But surely A may induce a false belief that p in B, and thus deceive him in this sense, without knowing, or even believing, that p is false. Indeed, A may believe that p is true, and he may have intended to communicate this to B by telling him that p. This is not a case of intentional deceiving, but it is a case of deceiving nonetheless.

Mele appeals to the dictionary to justify his use of ‘deception’. According to The Concise Oxford Dictionary, to be deceived, one must be mistaken or deluded, and to deceive, one must make someone believe what is false, mislead purposely, or be unfaithful to, or (archaic) disappoint. Thus, to be deceived about X is to be in error with respect to X. It may be safe to appeal to a reliable dictionary in order to solve a verbal problem. However, there are some problems with taking a lexical approach to the attempt to clarify a concept with philosophical significance. One obvious problem is that lexicographers only tell us how a term has been used over some time by native speakers. They do not tell us which sense of the term is philosophically interesting or relevant, or whether it is self-consistent. For instance, people might use the term ‘knowing’ in the sense of believing, but we as philosophers would not like to have our theory of knowledge based on this sense of the term. Hence, a better way to understand the notion of deception is to see how the notion is understood by different

---

philosophers.

Mele's notion of deception is not universally shared by other philosophers. McLaughlin holds that a deceiver is typically an intentional misleader. Donald Davidson also points out that deception must be intentional. Although self-deception is not lying to oneself, the self-deceiver must intend his deception so that it will not be confused with self-induced deception, such as when one does something intentionally with the consequence that one is deceived. According to Davidson, self-deception must satisfy the following conditions: 1) $S$ believes that data $d$ supports that $p$; 2) That $p$ motivates $S$ intentionally to manipulate data $d$ in such a way that it causes the belief that $\neg p$; 3) $S$ believes that $\neg p$ as a result of his manipulation of $d$.\(^3\)

Psychologist Paul Ekman, a known expert on deception, defines deception as follows:

(0)ne person intends to mislead another, doing so deliberately, without prior notification of this purpose, and without having been explicitly asked to do so by the target. There are two primary ways to lie: to conceal and to falsify. In concealing, the liar withholds some information without actually saying anything untrue. In falsifying, an additional step is taken. Not only does the liar withhold true information, but he presents false information as if it were true. Often it is necessary to combine concealing and falsifying to pull off the deceit, but sometimes a liar can get away just with concealment.\(^4\)

Lack of a consensus on the notion of deception among philosophers and experts means the essence of deception is hard to capture. This is why, I think, Mele has to appeal to the dictionary to settle the concept. However, what is problematic in his approach to defining the notion of deception is that he takes one of many criteria from the notion of other-deception while leaving other criteria out. That is, on the one hand, he insists on falsehood as the necessary condition for self-deception. On the other hand, he does not believe that the deceiver

\(^3\)Davidson, 1985, p.144.

\(^4\)Ekman, 1985, p.28.
need either believe that the opposite belief is true, or have the intention to deceive. This arbitrary choice of criterion seems to me to be motivated; i.e., this way of defining deception is convenient for Mele to solve the paradoxes of self-deception.

1.2 To err is human. I would argue that deception entails error but the other way around is not the case. Mele holds that "When we say that a person is deceived about something, X, we often mean simply that he is in error with respect to X, or that he holds a false belief concerning X." He is clearly equating deception with having a false belief. Moreover, deception is typically done by A's inducing a false belief that p in B, but unintentionally inducing a false belief in someone is just unintentional misleading. To be deceived with respect to X is not the same as, out of ignorance, to be induced to error with respect to X, as long as the ignorance is not intentionally created by oneself or by others.

Deception entails misleading, but it is not the case the other way around. Misleading can take place when error occurs during communication. Some cases of misleading are not in themselves forms of deception. As a little child, I was misled by my grandmother into believing that the earth is square and heaven is round. However, I cannot accuse my grandmother of deceiving me because I believe that she sincerely believed what she told me was true till her last day, and there is no reason for me to think that she concealed or falsified what she took to be true when she told me about the earth and heaven. Such cases of misleading happen all the time and, for the reason I stated earlier, I do not think it is reasonable to say this is a form of deception. Another unintentional misleader, say, an over optimistic or over pessimistic person may not be a self-deceiver. A person who tends to be misled by certain attitudes and thereby makes less than accurate judgments about his situation is not necessarily a
self-deceiver unless he tries to conceal or falsify the belief that he does not want to be the case.

The expression that the appearance of the world is deceiving should be looked at as an extended usage of the term ‘deceiving’. One might argue that this extended usage of ‘deception’ is the most appropriate usage in self-deception. The notion of deception that Mele adopts may be treated as a special case. This special sense of deception may serve as an explanation of self-deception, which may be a species of misleading. Real deception only occurs interpersonally. This line of thought cannot be accepted without further argument. I do not believe that we need to pick a special sense of deception in order to explain self-deception. If we take self-deception to be a form of deception, then the difference between other-deception and self-deception lies not in the notion of deception but, rather, in the notions and problems associated with self and others. Even though Mele’s notion of deception may help him resolve the paradoxes of self-deception, the problems that he has successfully solved may not be those of real self-deception since the notion of deception in his account has been heavily watered down. However, rejecting Mele’s watered down notion of self-deception does not mean that we have a real paradox of self-deception. Allowing the agent to have some level of perception that the belief \( \sim p \) is true should not make it paradoxical for the agent to believe that \( p \).

1.3 Mele has provided no ground for deception. Deception presupposes truth. But the ground for conveying or distorting truth is something else. I would argue that deception is possible only if games in general and language games in particular are possible. It is on such a ground that truth and falsehood are meaningfully understood and the rules of the game can be manipulated. Only then can we hold a deceiver culpable when the rules are violated. If I am right, then we might say that deception is the product of language games, when the word
'language' is understood as means of communication.

A rock cannot deceive, but arguably an animal might. Unless I am talking metaphorically about deception, it will be very odd for me to tell you that I was once deceived by a rock. I can play with rocks but rocks do not play games with me. On the other hand, it is not unusual for a dog trainer to say to someone that her dog deceived her at some point. This is, there is a big difference, to an experienced dog handler, between error and deception displayed in a dog’s behaviour. Philosopher and dog trainer Vicki Hearne has made an interesting observation of her dog Salty’s lying:

Retrieving makes possible a new sort of truth between Salty and me. It also makes new sorts of deception possible, and, in fact, the new truth is possible against the possibility of the new deception. The day comes when I lay the dumbbell on the ground, move about a leash length away from it, stand with Salty sitting at my side, the two of us facing the dumbbell, and, without my hand on the dumbbell, send my dog. She leaps out eagerly, with every appearance of sincerity, swiftly around in all directions industriously as if to say, “I’d love to bring it to you, boss, but I just can’t find it!” It is unlikely that she doesn’t know where it is, so I get on her ear and correct her, and she screeches with the sting and indignation of it. Suddenly she “remembers” that it’s under her chest and picks it up ... Soon, Salty explores another form of dishonesty, whose syntax is more like the syntax we have in mind when we think of lying as opposed to deception or evasion. Now I am throwing the dumbbell out as far as twenty or thirty feet, and it lands between Touchstone and a stick from the pepper tree. Salty picks up the stick rather than the dumbbell. This is almost like saying, “It’s not this one but that one,” although there are some queer things about it. For one thing, I assume that Salty doesn’t imagine that I don’t know the difference between the stick and the dumbbell. Salty is lying about herself, not about the dumbbell. She wants me to believe that she believes I meant the stick rather than the dumbbell ... She looks a bit sullen when I don’t accept these “retrieves.” It is as if the forms are so deeply if inarticulately felt for her that she must test my commitment to them thoroughly ... that she is playing around with the story I am trying to tell, is curious enough about it to be willing to risk the corrections.5

One can dispute with Hearne about her interpretation of Salty’s behaviour, but the point is not whether Salty is really lying or not; rather, on what ground can deception be attributed

5Hearne, 1982, pp.72-73.
What is the minimum criterion according to which deception is possible or attributable? It seems to me that it must lie in communication. We cannot attribute deception to a rock because rocks, in ordinary language, can communicate neither with us nor with each other. On the other hand, if we believe that there is some communicative relationship between rocks, it is not hard to imagine one rock deceiving another. For example, many of us do hold the belief that biological creatures do communicate with the rest of the eco-system; hence people (such as farmers and gardeners) can say that flowering trees are fooled by the weather when they bloom at the wrong time of the year due to unseasonable weather. This way of using the expression ‘being fooled by’ may sound metaphorical. However, it is not hard to imagine that some people might treat the expression literally if they also literally take nature to be an intentional system.

Although deception does not necessarily have to be intentional it must include the motivation that leads the agent to conceal the truth. The deceiver who hides something from his potential victim does so for a reason. The deceiver does not want the victim to face a certain truth, but this want alone is not sufficient for the deceiver to deceive. Sometimes, I can simply withhold information from someone so that the person does not reach truth, say the PIN number of my debit card. Only when I believe that what I want is likely to be threatened, that is, what I want may not turn out to be the case, then the need to make an effort to hide the relevant truth might arise. Although the goal of deception is to satisfy the pertinent desire, it may take more than a pertinent desire to motivate one to adopt a deceptive strategy in order to achieve one’s desire. This is what Mele has overlooked in his account.

So far, we have discussed a number of similarities between self-deception and other-deception. Obviously, self-deception is different from other-deception. For instance, in self-
deception, the deceiver and the victim are the same person while other-deception will involve at least two people. Moreover, the victim's believing that \( p \) in other-deception might be warranted, but the self-deceived belief is usually unwarranted. Charles Daniels argues that it is not appropriate to explain self-deception by modelling it on interpersonal deception. First, self-deception is irrational but other-deception does not have to be so. Second, in self-deception, when \( A \) says wrongly that \( p \) to \( B \), \( A \) may warrant \( B \) to believe that \( p \), but \( A \)'s saying wrongly that \( p \) to himself does not warrant him to believe that \( p \). Moreover, some characteristics are true in the cases of self-deception but not in the interpersonal deception. First, to be deceiving oneself, one may simply refuse to believe a certain proposition, *irrationally discounting* the evidence that objectively warrants one to believe it. Second, in self-deception, due to certain emotions, the self-deceiving person, over a period of time, seems to *magnify the evidential weight of* certain relevant evidence he has. Finally, the self-deceiving person is not willing to put his belief (or lack of it) to the test of other's experiences and intellect.\(^6\)

One might argue that other-deception is more offensive than that of self-deception. I would argue that this may be so in the sense that only one person will be the victim of self-deception but more than one person will be victims in other-deception. Even so, the severity of the offence ought to be measured by the attempt of the act of deception as well as the consequence of the deceptive act. Looked at this way, it is not so clear that other-deception is always more offensive than self-deception. One might also argue that the self-deceived person knows the reason neither for his believing that \( p \) nor for carrying out his self-deceptive project. That is, the self-deceiver does not know what he is doing, but the other-deceiver usually knows

\(^6\)Daniels, 1974, pp.248-252.
about his deceptive intention and strategies all along. However, other-deceivers can also be wrong about their motivation for deceiving someone else. If I am right to say that self-deception is a form of deception, then we should not abandon the full notion of deception in order to accommodate the philosophical explanation of self-deception. Instead, we should focus on whether the deception is reflexive or not and how the routes by which the deceived beliefs are formed differ between self-deception and other-deception.

1.4 The issue of whether deception necessarily involves a false belief is a tricky one. Mele’s Condition 1) in his list of characteristic and jointly sufficient conditions of a central case of one’s entering self-deception in acquiring the belief that \( p \) says that: The belief that \( p \) which \( S \) acquires is false. Mele argues that Condition 1) is purely lexical in that, by definition, one is deceived in believing that \( p \) only if \( p \) is false. However, falsehoods sometimes cancel each other. One can start with a set of false beliefs that \( q \), that \( r \), that \( s \), etc. Then, through biassed thinking one ends up with a true belief that \( p \). Mele’s response would be that the agent may be self-deceived in believing that \( q \), that \( r \), that \( s \), etc., but he is not self-deceived in acquiring the belief that \( p \), if the belief that \( p \) turns out to be true. If the belief that \( p \) is true, one cannot be said to be self-deceived in believing it, even though the agent is self-deceived. That is, the agent’s believing that \( p \) is believed on the basis of his self-deceived believing that \( q, r, s \).

Cognitive scientists tell us that our brain has evolved in such a way that forming

---

\(^7\)Dr. Catharine Talmage pointed this out to me during one of our many discussions about my present work.

\(^8\)Mele, 1997a, p.95.

erroneous beliefs is generally inevitable. We are, as it were, having false beliefs by default. If cognitive scientists are right, and if we follow Mele's notion of deception, then we are all self-deceivers by nature. Even if we could confidently reject the belief formation story given by cognitive scientists, philosophers have already realized for a long time that we are mistake-prone creatures. Errors in our belief formation may come from without, such as lack of sufficient information at the time of the belief formation, and from within, such as lack of imagination to conceive possible alternatives. Moreover, some evidence is inaccessible to us because we lack certain background knowledge against which it is necessary for the evidence to be perceived (i.e., data is theory-laden). Even if we switch on our critical thinking tool and are in company with the most critical minds all the time, self-deception may still be unavoidable, if we accept Mele's account.

I do not believe that the lexical approach is the best one, but I do not deny that the self-deceived belief is usually false. My objection to Mele's Condition 1) is that it greatly limits our ability to attribute self-deception to a person or to oneself. We may find cases that are considered garden-varieties of self-deception and be unable to call them such because we cannot gain access to the fact correspondent to the belief that the person deceptively held. For instance, the evidence suggests that the belief that \(~p\) is likely to be true, but \(S\) desires that \(p\) and, by treating the evidence in a biassed way, \(S\) gets himself to believe that \(p\). Even so, the belief that \(p\) is something that might just be true without our ever knowing it. Much theoretical knowledge and empirical knowledge has this nature. We can also use one of the "stock examples" of self-deception that Mele gives to illustrate this point:

Sid is very fond of Roz, a college classmate with whom he often studies. Wanting it to be true that Roz loves him, he may interpret her refusing to date him and her reminding him that she has a steady boyfriend as an effort on her part to “play hard to get” in order to encourage Sid to continue to pursue her and prove that his love for her approximates hers for him. As Sid interprets Roz’s behaviour, not only does it fail to count against the hypothesis that she loves him, it is evidence for the truth of that hypothesis.11

On first looking at this story, we might conclude that Sid is self-deceived into believing that Roz loves him. I believe that this conclusion is what Mele wants us to draw from the story that he has so told. Let us call the belief that Roz loves Sid the belief that $p$. Now, let us suppose that Sid and Roz are not some characters that Mele made up to match his theory, but that they are real people who we often run into. Are we sure that Sid’s believing that Roz loves him (that $p$) is false? Well, the story can be told differently. Upon our further digging into their story, we might find out that Roz does in fact love Sid. Call this the story of Sid and Roz (2) if you wish. Even if no one knows this, the belief that $p$ is true may still be a fact. It is possible that, even if everything else in their story remains the same, the belief that $p$ is true. Love is a very complicated thing. Roz may love Sid even though she also means it when she refuses to date Sid and she means it when she reminds him that she has a steady boyfriend, etc.

Sid may not believe that ‘no’ usually means ‘no’. If so, then Sid is wrong about every interpretation he has made about Roz’s behaviour except one thing, he is right about the very desired belief that $p$. Even so, Sid has adopted some self-deceptive strategies (such as positive misinterpretation) to sustain the belief he wants. Since the belief that $p$ is true and the belief that $q$ (the belief that Roz is playing hard to get) is false, Sid is self-deceived with respect to the belief that $q$ rather than $p$.

---

11Mele, 1997a, p.94.
Again, if we can imagine that the belief that $p$ is true, then why can’t we imagine the belief $q$ also being true? It is possible that not only does Roz love Sid, but also that Sid has made a true interpretation of Roz’s behaviour. Even though they are not warranted, the beliefs that Sid has about Roz may turn out to be true. Therefore, according to Mele, he is not self-deceived at all. In real life, we would have great difficulty finding the truth with regard to Sid’s belief about Roz’s love. It is no use to say that the truth lies in Roz’s own account because Roz may or may not know the truth with respect to her own behaviour and whether or not she loves Sid, since it is possible that Roz is not in touch with her own feelings and that she is unaware of the nature of her own actions when she is with Sid. If such skeptical doubt has any legitimacy, then we shall never be confident in attributing self-deception to Sid. Most of all, how do I, as a self-deceived person, find out that I have deceived myself? Whether anyone is self-deceived or not will be determined by the facts to which his beliefs correspond. Self-deception, looked in this way, seems to be a matter of luck, or lack of it. If it is so difficult to attribute self-deception to a person by Mele’s account, then Mele’s account of self-deception seems lacking in applicability. The empirical applicability of the theory may not be Mele’s primary concern, letting the state of affairs have the final say as to whether one is self-deceived remains a problem nevertheless.

I would argue that $S$ is self-deceived with respect to $p$ if $S$ is motivated to manipulate data relevant to the truth value of the belief that $p$, even if $S$ ends up being less than 50% certain with respect to $p$, even if $S$ merely guesses that $p$. Believing is a special cognitive attitude. Béla Szabados holds that the criteria for full blooded believing should be psychological (having a feeling of conviction that $p$), behavioural (being disposed to act as if $p$ is true) and logical
(producing reasons for $p$ when needed).\textsuperscript{12} I think that it can be quite rational for one to believe that $p$ and at the same time believe that $\neg p$ may be true. On balancing the evidence relevant to the belief that $p$ and other non-evidential considerations, one may believe that $p$ because one judges that it is better to believe that $p$ than to believe that $\neg p$. Even so, one is aware, given the obvious counter evidence, that $\neg p$ may be true.

We need to decide which should be counted more when it comes to identifying a self-deceiver: should $S$’s thinking process itself be more important, or should it be the degree of certainty that $S$ has about $p$? I believe that the former should have more weight. Even though we cannot be self-deceived into believing that $p$ without believing that $p$, we do not have to believe that $p$ in order to be self-deceived \textit{with respect to $p$}. If you hope that $p$ as the result of my attempt to mislead you to believe that $p$, I have committed deception with respect to $p$ even though, as the result of my attempt, you do not believe, but only hope, that $p$. Attitudes other than beliefs may also serve the purpose, such as estimation, guessing, or even hope, provided that they are the result of self-deceptive thinking. A deceiver need not be omniscient or omnipotent. The failure of a deceiver (or a self-deceiver) ought to be taken as a sign of his lack of skill or luck rather than as a lack of deceptive attempt.

Finally, if one must believe that the likelihood that $p$ is greater than 50\% in order to be self-deceived, then a person, say, a skeptic who has a lot of doubts but who is not certain about anything will be immune from self-deceptions. I do not know if being such a skeptic is possible, but suppose that it is so. Suppose such a skeptic’s thinking is as biassed as you can imagine and he is as deceptive as anyone can be. Yet it is, by Mele’s definition, impossible for

\textsuperscript{12}Szabados, 1985, p.157.
him to be deceived, simply because he is not more than 50% certain about anything. It seems to me implausible to make the falsehood of a belief that $p$ a necessary condition for deception or self-deception with respect to that $p$.

Mele asks, “What must be added to false belief to yield a condition of self-deception?” The answer he gives us is in the second and third conditions. In the following sections, I will deal with each condition in some detail.

2. Desire

In this section, I will be focussing on the notion of desire. The notion of desire is extremely important in philosophy, especially in philosophy of action and philosophy of mind. It is also important to those who are interested in the problem of self-deception. Self-deception involves behaviour that is driven by certain states of mind. I do not attempt to come up with a complete account of desire as it is too complex for the scope of my present work. However, since I must discuss the notion of desire to a certain degree in order to show what I believe to be problematic in Mele’s account of self-deception, especially his Conditions 2) and 3), I shall first review some basic issues about desire.

2.1 The second and the third conditions in Mele’s list of characteristic and jointly sufficient conditions of self-deception of the belief that $p$ are the following: 2) $S$’s desiring that $p$ leads $S$ to manipulate data relevant to the truth value of $p$; 3) This manipulation is a cause of $S$’s acquiring the belief that $p$.

What is desire? Spinoza says that desire is “appetite accompanied by the consciousness
thereof' and that "we judge a thing to be good because we endeavour ... and desire it." Mele does not explicitly deal with the nature of desire in his works on self-deception. He merely states that one's want or desire is key to one's biassed thinking and self-deception. By motivation, Mele simply means want or desire, as he puts it in his notes:

In this book, I shall use the nouns 'want' and 'desire' in a broad, semitechnical sense that has considerable currency in the philosophical literature on action. As I shall use these terms, to say that S has a desire or a want to A is to say that he has some motivation to A, the propositional content of which makes essential reference to his (prospective) A-ing. I shall use the verbs 'want' and 'desire' in a correspondingly broad sense.14

In his *Springs of Action – Understanding Intentional Behaviour*, a book that came out three years after his *Irrationality*, Mele made some further distinctions among different types of wants or desires in order to capture different senses of the usage in standard English. He points out, for instance, that there is *occurrent want* and *standing/latent* wants. Roughly speaking, a standing want is one which aims at a long term goal and has a disposition to have occurrent wants for that goal under a range of conditions. It is a dispositional state because it is a disposition to acquire the occurrent wants. Neither type of wants needs to occur consciously.

2.2 What is the difference between a desire and a belief? There is at least one difference, namely, that desire is about goodness (or value) and belief is about being true. There are other ways to capture the differences. For instance, some philosophers hold that desires have a world-to-mind direction of fit and beliefs have a mind-to-world direction of fit, and that desire is about the future and belief is about present or past. Dancy, for instance, holds that, with

---


respect to being the elements of a motivating state, belief and desire can be distinguished in
cognitivist terms such that one represents the world as it now is, and the other represents the
world as it will be when and if the action is successfully completed. He argues that these two
representations are both necessary because, without the first representation, one has no idea of
what one is working from; without the second representation, not only one has no idea of what
circumstances in which one is acting, but also one has no idea what one is working to achieve.\(^\text{15}\)

Wayne A. Davis holds that the word ‘desire’ is ambiguous. It expresses two logically
independent propositional attitudes. One is called “volitive” desire and the other is called
“appetitive” desire.

While both tend to generate action, to spread from ends to means, and to yield pleasure
when satisfied, there are major differences: appetitive desire influences volitive desire,
not vice versa; volitive desire a better index of action, appetitive desire of pleasure; and
only volitive desires are based on reason, influenced by value judgements, or entailed by
intentions.\(^\text{16}\)

Annette Baier argues that desire is not reducible to other mental phenomena such as
motivation, intention, or seeing good reasons. She says:

Desire is of most interest as a psychical phenomenon when it, with its variant of
intentionality, is seen in relation to all its mental relatives with their variants, to
pleasure, to satisfaction, to love, to beliefs about the good and how to get and sustain it,
to discontent, to hopes and fears for the future, to confidence and to lack of confidence,
to a sense of ability and to ignorance to the extent of one’s ability; to depression, grief,
despair, homesickness, nostalgia; to longing, craving, lusting, itching, wanting,
preferring, and intending. Desire is what it is, and not these other things, and they are
what they are, and not desire. It will be understood best, philosophically, when it is
seen in clear relation to all of its many relatives among the actions, passions, and the
mixed active/passive states of our complicated, variegated, self-complicating and self-

\(^{15}\)Dancy, 1993, pp.3, 13-14.

\(^{16}\)Davis, 1986, p.63.
diversifying souls.\textsuperscript{17}

Maybe we need to introduce a broader notion than that of desire or want. Davidson uses pro attitude to cover what an agent wanted, desire, prized, etc.\textsuperscript{18} This seems to be consistent with Mele’s account of self-deception. To keep things a bit simpler, when I use the word “desire” in the rest of my thesis, I mean pro attitude.

3. \textit{Motivation}

In this section, I will discuss Mele’s Condition 2) and Condition 3) in that order. I will then discuss why I think that Mele’s account is inadequate. I will argue that the reason explanation (i.e., the explanation is given within ‘the space of reasons’) is just as important as the causal explanation (i.e., the explanation is given within the realm of psychology and physics). I will then try to work my way towards a reconciliation with Mele’s account of motivational strength. In so doing, I hope to prepare for the introduction of my own reason account of self-deception, which will be in the final section of this chapter (section 4).

3.1 The Condition 2) -- $S$’s desiring that $p$ leads $S$ to manipulate data relevant to the truth value of $p$ -- cannot be treated literally. There is no reason for us to believe that one can be self-deceived into believing that $q$ (if $q$ is false) while desiring that $\neg q$, as is the case in our story of Sid and Roz (2). However, Sid need not desire that $q$ (that Roz is playing hard to get) in order to believe that $q$, since it is possible that Sid wants Roz to be straightforward with him and not to play the hard to get game with him. Suppose Sid wants it not to be the case that Roz is

\textsuperscript{17}Baier, 1986, p.59.

\textsuperscript{18}Davidson, 1980, pp.3-4, 6, 11, 83-87.
playing hard to get, yet he gets himself to believe that this is the case; does this mean that his want that \( \neg q \) motivates him to believe that \( q \)? It may be tempting to say that a con attitude, such as fear or hatred, may motivate one to believe what one fears or hates to be the case. One might argue that paranoiacs or hypochondriacs are often deceiving themselves into believing the things that they fear the most so it seems to be the case that one’s con attitude, just like one’s pro attitude, may lead one to be self-deceived. \( S \) might be self-deceived into believing that he has an incurable disease even though there is no medical evidence of that at all. One might say that \( S \)’s self-deception has something to do with his fear of having such a disease and so that he is motivated by the fear, rather than the desire of such a disease. This may be how we usually explain the typical behaviour of a hypochondriac. However, I do not think that hypochondria, jealousy or paranoia are cases of self-deception. In these cases, the victim is so overwhelmed by his fear that his perceptions have altered quite dramatically. A hypochondriac’s perception of his physical suffering is every bit as real to his as the kind of pain caused by real disease. Such feeling itself can be a convincing form of evidence, for him, for believing that he is suffering from some serious illness. Misperception of this kind is not self-induced by one’s own intentional action. Thus it will not be fair to say that hypochondria or paranoia are cases of self-deception.

There are cases of motivated false believing that are not clearly cases of self-deception. One may believe that to give a certain amount of thought to unpleasant things may prevent these things from happening, or at least, one is better prepared for the worst by thinking about the worst scenario. Over time, due to selective focussing, selective data collecting, misinterpretation, etc., one gets oneself to believe the very thing that one least wants. This thinking process may be full of pseudorationality, yet it may not involve intention to
manipulate data relevant to the belief held in the same way as in self-deception. The biassed thinking process is not motivated by one’s perception that evidence counts against the belief held. It is motivated by a different kind of belief, something like ‘thinking makes it not so’. What is crucial is that the biassed thinking does not involve the intention to conceal or falsify the truth value of the belief opposite the one held. If it does, then I would take it to be a form of self-deception, assuming that it is possible for one to be self-deceived by one’s con attitudes as well as one’s desires.

What motivates Sid to believe that $q$? Mele’s answer would be this: Sid need not have the desire that $q$ in order to be motivated to believe that $q$. Sid may be motivated (by the desire that $p$) to believe that $q$ even though he has no motivation to believe that $q$. Mele made a similar argument about a different case. He argues that $S$ wants to empty his bathtub by throwing the water through the bathroom window. If he does so he will fill up a wading pool beneath the window. He does not want to fill up the wading pool, but, in order to empty the tub, he throws the water through the window, thereby filling the pool.

Nevertheless, the initiation of $S$’s filling the pool is not motivation independent. $S$ filled the pool by dumping the water out the window, and he was motivated to do the latter. $S$’s filling the pool, we might say, was motivated by his desire to empty the tub. And, in this sense of the quoted phrase, we may plausibly say that $S$ ‘was motivated to fill the pool’ even though he had no motivation (desire) to fill it.19

This answer still does not tell us why one’s desire that $p$ motivates one’s belief that $q$ rather than the belief that $\neg q$. After all, if Sid desires that $\neg q$ (the desire that Roz is not playing hard to get), the belief that $\neg q$ ought to have more hedonic quality to him and therefore he should be more likely to believe it than to believe its opposite. I believe that the answer can be found

---

under a different light. Sid's believing that $q$ is not motivated by his con attitude towards that $q$. I would argue that Sid cannot bear the inconsistency of co-existence of the beliefs that $p$ and that $\neg q$ in his belief system. He must accept either one or neither of the two beliefs at one time. On balancing his desire and belief, Sid chooses to reject the belief that $\neg q$ in order to keep the belief that $p$. If one's intention to believe that $q$ does not entail the desire that $q$, then Sid does not believe that $q$ because of a desire that $q$ but, rather, he believes that $q$ for the sake of maintaining the consistency with the other belief, the belief that $p$, that he intends to hold. In other words, the demand for consistency in one's belief system can motivate one to believe something other than what one desires.

3.2 I said in the preceding paragraph that, on balancing his desire and belief, Sid chooses to reject the belief that $\neg q$ in order to keep the belief that $p$. But what role does the desire that $\neg q$ play in Sid's believing that $q$ and rejecting the belief that $\neg q$? Sid's desire that $\neg q$ does not seem to motivate him to adopt any of the strategies of self-deception, such as negative or positive misinterpretations, which might cause him to believe that $\neg q$. Sid's desiring that $\neg q$ may play no motivational role in Sid's self-deceived belief that $q$ at all. Why doesn't it? How is the balance that ends up in favour of the belief that $q$ reached in Sid? Some might argue that the fact that Sid chooses or wants to believe that $q$, rather than to that $\neg q$, shows that he is more strongly motivated to believe that $q$ than to believe that $\neg q$. In other words, motivational strength equals the power which causes one's mental or physical states. There is no doubt that the issue of motivational strength is relevant, but such an answer adds no further information to what we already know, as using the term 'motivational strength' in this way is merely to add an empty term to the problem. We need an account which explains where a desire gets its
motivational strength from and where it leads to. Here, I am in favour of Mele’s account of the issue. Mele argues that the motivational strength of a state should not be simply a measure of its capacity to produce appropriate intentional behaviour. Otherwise, all it amounts to is that “wants move us to act in virtue of their capacity to move us to act”, rather than offering the features of wants, and that, Mele believes, goes into a vicious circularity. Moreover, Mele argues that, when an agent has seemingly conflicting desires $A$ and $B$, he may deliberate on what he wants (the most) to do (i.e., desire) and on what to do (i.e., intention). He may want the most to do $A$ yet he intends to do $B$. Consequentially, he does $B$.20

Mele further argues that when $S$ has a number of competing desires, the motivational strength of each desire varies. The motivational strength of a desire cannot be determined by the felt intensity of the desire, if our intentional behaviour is to accord uniformly with our strongest desires: sometimes, our most affectively intense desires are not the desires on which we act. Nor is desire-strength universally fixed by agents’ assessments or evaluations of the objects of the desires, as garden-variety instances of akratic action show.21 That is, the motivational strengths of the intention to do $A$ and the desire for $B$ are not always the same for an agent at one particular time. Mele further argues that the causal relation and the strengthening relation among different desires work in their own different ways. He has introduced a technical term, the positive motivational base of a desire, to explain the different kinds of relationships among desires. The positive motivational base “is the collection of all occurrent motivations of the agent that make a positive contribution to the motivational strength of that desire.” However, “not every occurrent motivation that figures in the

---


21Mele, 1992, p.76.
causation of a particular desire is in the positive motivational base of that desire."^{22} This is how Mele's argument can help to deal with the question of how the balance that ends up in favour of \( q \) is reached in Sid. Among many desires that Sid has are his desires that \( p \) and that \( \sim q \). Now let us imagine that Sid is deliberating at this moment on how to achieve his desire that \( p \). Sid comes to a (false) belief that \( p \) implies \( q \). Consequently, he forms a second-order desire that he believes that \( q \) (\( D^2Bq \)) while still retaining his first-order desire that \( \sim q \). Now, Sid starts to deliberate about how to achieve this second-order \( D^2Bq \) and he comes to believe that doing \( A \) (for example, collecting supporting evidence for the belief that \( q \)) is the best way to achieve his \( D^2Bq \). It is worth noting that \( D^2Bq \), although being a causal antecedent to \( A \)-ing, does not strengthen Sid's desire to do \( A \) (\( DA \)), for the strength of the desire to do \( A \) may be wholly derivative from the strength of Sid's desire for that \( p \) (\( Dp \)).

Let me summarize what we just discussed ("D" stands for "the desire that ..."); "D"' stands for "the second-order desire that ..."); "B" stands for "the belief that ..."):

\[
\begin{align*}
D(p \& \sim q) & \land B(p \rightarrow q), \\
Dp & \rightarrow D^2Bq, \\
B(A \rightarrow D^2Bq), \\
\therefore DA, \\
\therefore A\text{-ing}, \\
\therefore D^2Bq, \\
\therefore Bq\text{-ing}.
\end{align*}
\]

Mele argues that, even though the causal relation and the positive motivational base can

---

^{22}\text{Mele, 1987, p.67.}
be functioning separately, "the positive motivational base of any intrinsic desire is at least partly internal to the desire itself."[^23] That is, a desire that is the source of its own motivational strength may also derive motivational strength from other desires. Sid's desiring that \( p \) may also derive motivational strength from the rest of his desires in the causal chain, such as, \( D^2Bq \), \( DA \), etc.

Conversely, an occurrent motivation may strengthen a desire without having a causal relation to the formation or retention of that desire. Having already acquired the desire to do \( A \) (i.e., to manipulate the truth value relevant to the belief that \( q \)), Sid discovers that \( A \)-ing will contribute to the satisfaction of his desire for \( B \) (such as meeting Roz's former roommates and her female friends). If Sid's desiring \( B \) strengthens his desire to do \( A \), his desire for \( B \) will be in the positive motivational base of the desire to do \( A \) without causally contributing to Sid's continuing to desire to do \( A \). The formation and sustaining the desire to do \( A \) may be causally independent of his desire for \( B \).[^24]

\[
\begin{align*}
B(A \rightarrow B) & \land DB, \\
D(p \land \neg q) & \land B(p \rightarrow q), \\
Dp & \rightarrow D^2Bq, \\
B(A \rightarrow D^2Bq), \\
\therefore DA, (DA may be causally independent of DB). \\
\therefore A \text{-ing.}
\end{align*}
\]

It is not hard to imagine that Sid, at some point in his deliberation on how to achieve his desire that \( \neg q \), realizes that to achieve his desire that \( \neg q \) means that he cannot satisfy his desire

[^23]: Mele, 1987, p.68.
that $p$. In other words, he must choose between the satisfaction of the desire that $p$ and the satisfaction of the desire that $\neg q$. The desire that $p$ seems to be stronger to Sid than the desire that $\neg q$. This answer may not satisfactorily explain why Sid’s desire that $\neg q$ fails to motivate in formation of a relevant belief. However, the question at issue is ‘how is the balance that ends up in favour of the belief that $q$ reached in Sid?’ and it has been answered by Mele’s account of motivational strength.

These arguments of Mele’s about motivational strength seem to have a considerable explanatory power. However, I would argue that without Sid’s believing that the belief that $p$ is obtainable and that $\neg p$ may be the case, the self-deception in believing that $q$ might not occur. Simply having a desire that $p$ alone may not lead Sid to misinterpret Roz’s behaviour to the extent that he is self-deceived. It may only lead him to daydream about Roz’s affection towards him, or it may be Sid’s wishful thinking that Roz will one day love him. It seems that desire alone is insufficient to motivate, and that desire and belief are correlative. The interaction may be logical.

3.3 Mele says, “As I shall use these terms, to say that $S$ has a desire or a want to $A$ is to say that he has some motivation to $A$, the propositional content of which makes essential reference to his (prospective) $A$-ing. I shall use the verbs ‘want’ and ‘desire’ in a correspondingly broad sense.”

25 Given this how exactly does one’s desire motivate one into self-deception? Mele’s answer is a causal story. The causal mechanism combines motivated self-deceptive strategies (what Mele calls ‘hot’ mechanism) with the common sources of unmotivated irrational beliefs

---

(what Mele calls 'cold' mechanism). The 'hot' mechanism (i.e., the motivated self-deceptive strategies) can be further categorized into the internal-biassing strategies and the input-control strategies. Each can be further divided into positive and negative misinterpretation of data and selective focussing or attending to data. The 'cold' mechanism (namely, the common sources of unmotivated irrational beliefs) includes 1) The Vividness of Information, 2) The Availability Heuristic, 3) The Confirmation Bias, and 4) The Tendency to Search for Causal Explanations.26

These 'hot' mechanism (or the motivated deceptive strategies) and 'cold' mechanism (or the common sources of unmotivated irrational beliefs) are of particular importance in Mele's theory of self-deception because it is here that most of the explanatory power of Mele's theory resides. There are at least three questions that need to be answered here. 1) How can one's employing a deceptive strategy, X, result in one's believing that p? 2) How can one's desiring that p motivate one to employ a deceptive strategy X? 3) How can one's employment of a deceptive strategy X, motivated in the manner that it is, fail to be self-defeating?

Mele's answer to the first question is that self-deceptive strategies exploit the common sources of irrational beliefs without intentional guidance for their effectiveness. In other words, 'hot' mechanism (motivated deceptive strategies) may promote 'cold' mechanism (the common sources of unmotivated irrational beliefs) without the guidance of the agent. His answer to the second question is that one's desire may motivate one to employ a deceptive strategy in a variety of ways. Mele's answer to the last question is that the strategy paradox is only real in cases of intentional self-deception.27

I would argue that, in his last answer, Mele misses the crucial possibility that these

27Mele, 1987, p.149.
strategies may be guided by the intention (which is part of the second conclusion) to conceal or falsify the truth value of the opposite belief (that \( \sim p \)). The intention to conceal or to falsify may come in very shortly before S’s manipulation of the truth value relevant to the belief that \( \sim p \) takes place, so that it may be both an antecedent and an efficient cause of S’s data manipulation. Moreover, the intention to conceal or falsify may also contribute to the sustaining of the data manipulation. I will discuss this issue in the final section of this chapter (section 4).

3.4 Mele’s second answer above, that one’s desire may motivate one to employ a deceptive strategy in a variety of ways, puzzles me the most. I have no intention to deny that desire does motivate. The problem lies in the way he treats desire and motivation. He takes desire as the only element that motivates. On the other hand, according to Mele, the basic strategy for self-deception is very much the same as the unmotivated irrational belief formation process, except that in the case of self-deception the agent is motivated by desires. In other words, it is the ‘cold’ mechanism (namely, the common sources of unmotivated irrational beliefs) “primed by motivation” that marks self-deception apart from other false beliefs. For instance, his desire that \( p \) may make the data supporting \( p \) appear more vivid to the agent, so that the supporting data catches his attention. Alternatively, the agent’s desire that \( p \) may affect the hedonic quality of thoughts about different data, so that he decides to focus on confirmatory data rather than disconfirmatory ones.\(^{28}\)

If the studies\(^{29}\) of social psychology can be relied on, then we tend to think in a biased way anyway, even without desire being involved in many cases. If Mele is right in claiming

\(^{28}\) Mele, 1987, 149.

that one’s desire may be a contributing cause of one’s biassed thinking, then we must know why one’s not desiring something (such as hate and fear, etc.) cannot also be a contributing cause of one’s biassed thinking. Why must we think that it is the desire rather than something else, such as beliefs, health, environment, etc., that leads one to thinking in these biassed ways? It is not that we cannot accept that desire must play an important causal role in self-deception, it is that desire in Mele’s account seems to explain so well, almost too well, all the problems that we have in self-deception. This should make us suspicious.

3.5 Mele’s Conditions 2) and 3) in effect put S in the dark. When he manipulates his data, S does not think that he is misinterpreting the data, or that he is selectively focussing or attending on specific data. He could have prevented such mistakes had he noticed what he was doing to himself. Mele insists that the mechanisms of self-deception are unguided by the agent. In a way, the more successful the story about how motivation can enhance the sources gets, the less convincing it is to argue that the self-deceived person has any control over self-deception that he commits. It is his desire, over whose existence S has no control, that makes him think in biassed ways and eventually allows him to be self-deceived. We want to question, if S is not merely a dummy in Mele’s account, then where is this person in self-deception? Is the whole project just a matter of psychological states and events that happen to an individual? Or is it something that an agent does to himself? Or is it something true under both descriptions?

I am not convinced by Mele that hot (or motivated) cases of biassed thinking are all that hot. The other side of the coin is, of course, that the ‘cold’ mechanism (the common sources of unmotivated irrational beliefs) may not be all that cold either. Having a close look at these
sources of unmotivated irrational beliefs, one might find that they are so tightly woven into our rationality that it is impossible to avoid them when we deliberate. We seem to be stuck in this situation. In some cases, our biased thinking is extremely pervasive and almost impossible to avoid. We all have desires and we are not clearly aware of all of them. We err often. In each error one makes, there may be desires and biased thinking to be found if we look hard enough. Therefore, each human error is likely to be a case of self-deception.

This leads us to a further question posed by Bas van Fraassen, namely, whether one can doubt one's most intimate stories that one tells about oneself without entirely sacrificing self-knowledge? van Fraassen argues that it is generally agreed that one can be justified to say that someone else is self-deceived in believing that $p$, or that one was self-deceived in believing that $p$, but one could not rationally say that one is now self-deceived in believing that $p$. If so, then the attribution of self-deception makes sense only from an exterior point of view, a point of view that one does not (yet) share. However, this brings us to a skeptical doubt: one would always be in a position in which one could doubt one's conclusions about oneself. Since doubt feeds doubt, taking the possibility of self-deception seriously may undermine all the stories we tell about ourselves; for instance, it would be difficult for a theorist to see that he is not self-deceived about his account of self-deception.\(^{30}\)

The sources of biased belief that Mele cites are a central part of our reasoning process. It seems that to acquire and retain beliefs through these sources is just to be rational, at least in part. The question seems to be, how can we not be self-deceived if we must believe at all? If, as Sartre said, we are damned to choose what to believe, and if we must imagine the

\(^{30}\)van Fraassen, 1988, pp.135, 140-141.
alternatives, then we are damned to be self-deceived.

I think the answer should be that an adequate explanation of self-deception must contain both descriptions in order to answer the why and how questions. The 'how' questions may be sufficiently answered by Mele, but the 'why' question requires us to put the agent back into his place in his self-deceived thinking and believing, which Mele has failed to do. I would argue that, if \( S \) intends to falsify the belief that \( \neg p \), then he cannot be said not to know what he is up to. He might not know the strategies that he employed for his falsification of a belief. He might also not know that what he intends to do is to falsify a belief. Yet, if he tries to reject a belief the supporting evidence for which he is aware of, he must reject this belief intentionally. If I am right so far, then I think I have put \( S \) in the active place where he should be in deceiving himself without an intention to deceive himself.

Mele holds that cognition -- however intense, complete, and so forth it may be -- will not initiate an intentional \( A \)-ing unless the agent is motivated to \( A \), and that there is nothing else to turn the trick either: intentional action is motivated action.\(^{31}\) This is a Humean account of motivation. According to Jonathan Dancy, Hume's belief/desire thesis is roughly this: Every complete motivating state is a combination of belief and desire. A belief without desire cannot motivate one to act, a desire without belief is impotent. Humean desire/belief motivating states are this: Desires are essentially (internally) motivating states. Beliefs are contingently (externally) motivating states that are able to motivate when keyed into a suitable desire.\(^{32}\) However, most of our desires don't have any phenomenology. Even if one can feel a strong desire, as Dancy rightly asks, could the desire be present without itself motivating while other


attitudes are motivating? That is, why is it not possible that the desire that $p$ merely happens with $S$’s believing that $p$, instead of motivating $S$’s data manipulation? This is a very difficult question for Mele. In a more recent journal article, instead saying that $S$’s desiring that $p$ leads $S$ to manipulate data relevant to the truth value of $p$, Mele now only claims that desire may contribute to the biased thinking. This is a much weaker claim that is not all that interesting. It is reasonable to assume that our beliefs and desires work in tandem in initiating our intentions to act as well as in motivating our actions. In our Sid and Roz story (2), for instance, why does Sid’s desiring that $\neg q$ fail to motivate him to believe that $\neg q$? Mele can answer this by arguing that the desire that $\neg q$ and the desire that $p$ have different motivational strengths. The latter desire may have a stronger motivational strength than the former one. How do we know the motivational strength of Sid’s desire that $p$ is stronger than the motivational strength of his desire that $\neg q$? It is a circular answer, as Mele rightly points out, to say that the fact that the desire that $p$ has succeeded in motivating and the desire that $\neg q$ has failed to motivate means that the former has a stronger motivational force. We might be able to work out a complex theory to measure the different motivational strengths among our desires, as Mele has tried to do. However, focussing on motivational strength is not the only way to answer the question at issue. A different answer seems to me equally good, if not better, as an explanation of why Sid’s desire that $\neg q$ fails to motivate him to believe that $\neg q$. It is this. We can look into Sid’s belief system as well as his desires. Sid believes that his desires that $p$ and that $\neg q$ cannot both be satisfied. He believes that believing $q$ is consistent with the belief that $p$, and he also believes that believing the latter is conducive to the satisfaction of his desire.

---

33Dancy, 1993, p.23.

34Mele, 1997b, p.132.
that \( p \). Sid’s desire that \( \sim q \) fails to motivate because, on balance among Sid’s desires and beliefs, it is not worth pursuing. The desire that \( \sim q \) is significant to Sid only if the desire that \( p \) is important to him. The opposite is not the case. Without the desire that Roz love him, Sid would be much less concerned whether Roz were playing hard to get or not.

3.6 Mele insists that the self-deceiver’s data-manipulation is motivated by his wanting \( p \); the motivatedness of his intentional misinterpretation does not depend on his believing that the data supports \( \sim p \). Even though believing that the data supports that \( \sim p \) may contribute to one’s motivation in data-manipulation, it is one’s pertinent desire that does the conative work in data-manipulation, intentional or not.\(^{35}\) Moreover, Mele argues that the self-deceived reason is not the belief that ‘datum \( d \) counts against \( p \)’, but rather \( S \)’s desiring that \( p \) and believing that \( A \)-ing may achieve that \( p \).\(^{36}\)

One might question, without the belief that ‘datum \( d \) counts against \( p \)’, how \( A \)-ing can be conceptually connected with the manipulation of datum \( d \). If \( S \)’s \( A \)-ing has nothing to do with the manipulation of datum \( d \), then his manipulation cannot be viewed as relevant to the truth value of the belief that \( p \). Hence \( S \) is not self-deceived. If \( S \)’s \( A \)-ing is related to the manipulation of datum \( d \), then I would argue that \( S \)’s desiring that \( p \) alone is insufficient to explain why he \( A \)-ed. \( S \)’s belief that ‘datum \( d \) counts against \( p \)’ must be also taken into account for the explanation of \( S \)’s choice of action. That is, there is no clear connection between \( S \)’s desire and his reason for manipulating datum \( d \). Datum \( d \) is connected to \( S \)’s other beliefs. \( S \)’s desiring that \( p \) may make datum \( d \) appear to him more significant than it would be in the

\(^{35}\)Mele, 1987, p.128.

absence of the desire that \( p \) only if \( S \) believes that datum \( d \) matters a lot to the belief that \( p \).

Therefore, both \( S \)'s desiring that \( p \) and \( S \)'s belief that ‘datum \( d \) counts against \( p \)’ must be counted as a motivation for \( S \)'s \( A \)-ing.

I can think of at least one reason why Mele does not take it into account. Mele wants to solve the doxastic and strategy paradoxes of self-deception. His strategy for solving the paradoxes depends on the notion that deception is possible without the intention to deceive. A non-intentional deceiver is a person who is ignorant of the truth with respect to the belief in question. I would argue that, first, a non-intentional deceiver may be ignorant of the truth value of the belief that \( p \). He may falsely believe that \( p \) is true; but if he does not at some level perceive that the belief that \( \neg p \) is true, he is not a deceiver.

Second, in some cases, the agent’s awareness of the evidence against the self-deceived belief seems to be crucial to sustaining the self-deception. \( S \) wants \( p \). The belief that \( p \) was true in the past but now is false because things have changed over time. Suppose that \( S \) still wants \( p \) to be true and still believes it contrary to the available evidence. Now, if \( S \) has no idea at all that there is mounting evidence to support the belief that \( \neg p \), we would say that \( S \) is ignorant in maintaining the belief that \( p \), and it is not a case of self-deception. However, if \( S \) obviously ignores or tries to explain away this evidence, then we can say that \( S \) is now deceiving himself in believing that \( p \).

Third, notice that \( S \) cannot ignore something that he is not aware of. As Annette Baier says, “Ignoring is the same art as concentrating, seen from the side of what gets excluded from one’s main focus of attention.”\(^{37}\) Ignoring can be used as a strategy of self-deception for data

\(^{37}\)Baier, 1996, pp.53-55.
manipulation. If so, then S’s believing that ‘datum d counts against p’ (Bsd) plays an important role in motivating S’s believing that p. Bsd (i.e., S’s believing that ‘datum d counts against p’) is a precondition for S’s data manipulation. If Bsd is necessary for S’s data manipulation, then Bsd causally contributes to S’s self-deception in a much more important way than Mele has made out.

Fourth, that Bsd also explains why S interprets the data in one biassed way rather than in some other way. For instance, because Sid believes that it may very well be the case that Roz loves her boyfriend but not him, Sid is more likely to try to explain away why Roz frequently tells him that she loves her boyfriend than to try to explain away many other things, such as why Roz frequently changes her sneakers.

Finally, one’s realization of one’s inconsistency of beliefs demands that one give an explanation in order to gain coherence in a way that one’s realization of the conflict of desires and beliefs does not. The following cases show this: (1) S believes that p, that q, and that (q → ~p). (2) S desires that p, and S believes that q and that (q → ~p). In case (1), there is a clear urgency for S to find an explanation that can sort out the inconsistency among his beliefs. In case (2), there is no such urgency. S may be frustrated by realising that his desire that p may not be likely to be satisfied, but there is nothing so unusual about that. It is not clear to us why S’s desire that p could motivate him to conceal the truth value of the belief that ~p until we take into account that S believes that q and that (q → ~p). Hence, S’s believing that q and that (q → ~p) sets a departure point for his inappropriate manipulation of data relevant to the truth value of the belief that ~p.

I grant that it is not the opposite belief that motivates one to manipulate data. However, I would insist that awareness of the evidence that supports that ~p coexists with a self-deceived
belief that $p$. I agree with Mele that the self-deceiver need not be in a peculiar doxastic condition in believing both that $p$ and that $\neg p$. I also agree with him that one may not be able to show any case of self-deception that genuinely involves contradictory beliefs. However, what I want to emphasize is this: the self-deceived person's awareness of the evidence that supports that $\neg p$ must be there. He must, at some level, perceive that the opposite belief is true.

3.7 Mele does deal with the 'how' issue of motivatedness of desire, but his singling out of the desire that $p$ for the responsibility of entering self-deception is also problematic. We have a picture which shows us that desire plays an important role leading towards self-deception. However, Mele fails to explain why a person deceives himself. He does not give us a logical connection, in the sense of practical reasoning, between one's desire and one's employment of strategies and 'cold' mechanism. As I argued earlier, when the belief that $p$ is true and the belief that $q$ is false, if one desires that $p$ and that $\neg q$, then one can be motivated by the desire that $p$ and be self-deceived into believing that $q$ without being self-deceived in believing that $p$. The connection between the desire that $p$, and the false belief that $q$ is not accidental or brute. Given the other beliefs relevant to the desire that $p$ and the belief that $q$ that $S$ has, it is reasonable, from $S$'s point of view, for him to believe that $q$.

The connection between the desire that $p$ and the false belief that $q$ must be obtainable internally. That is, the answer to the question "Why do I deceive myself?" should result from reflecting on my own beliefs. The point here is not that I am the one who has direct access to my beliefs, wants, intention, etc., which are the only direct evidence for understanding the reason(s) for my actions. The question is, rather, how can I reasonably be expected to have access to the fact that I am adopting some self-deceptive strategies, such as misinterpreting
data? How can I reasonably be expected to be aware that some of my desires may enhance the ‘cold’ mechanism of irrational beliefs and lead me to self-deception? If I am not expected to be able to give my reason for the judgment I make about what to believe and what to do, then I am not being treated as a rational being. Being self-deceived is not an indication that I am incapable of being rational, any more than my occasional stumble and fall is an indication of my inability to walk. In contrast, the inability to explain one’s thoughts and intentional actions (for example, the manipulating of the truth value of a belief) by giving reason for them does cast serious doubt on one’s being rational.

I am trying to push the idea that we cannot explain self-deception adequately without showing that the agent plays an active role. We do a lot of things without any reason for doing them, even though we can give detailed causal explanations to each of them. I woke up early this morning and felt thirsty. I got up and went into the kitchen because I wanted a glass of water and I believed that I could get a glass of water there, but the first thing I did in the kitchen was step on my cat who was lying there resting. Why did I step on my cat? I can give a rough causal explanation for why and how I stepped on the cat. I am sure you can guess what went on and I will not bore you with all the details. The point is that, although I might be able to give you good causal explanation of why and how I did it, I cannot give you my reason for doing it because I had no reason for doing it at all. The reason for me to go to the kitchen was not so that I could step on my cat but, rather, to get a glass of water. Again, if the self-deceived person simply does not know what he is doing when he is deceiving himself, then he can simply say, when he is asked why he deceives himself, that he has no reason at all for doing so. Mele’s agentless causal explanation gets the self-deceiver off the hook too quickly.

Certain kinds of willfulness fit into Mele’s account of self-deception, but I consider it
otherwise. For instance, xenophobic people or people who advocate racial discrimination often want to keep certain false beliefs about the people they do not know and willfully turn a deaf ear to opposite opinions. Narrow-minded people often choose what they want to believe. They are not deceiving themselves but merely deprive themselves from seeing the truth.

4. An Explanation of Self-Deception

I accept Davidson’s concept of irrationality and argue that self-deception is not necessarily irrational. One is epistemically irresponsible to deceive oneself; however, if one does not accept, as one’s overriding principle of believing, that one should always believe what is best supported by evidence, then one is not irrational in believing something that is not warranted. Since self-deception is not necessarily irrational, we might ask, what is the reason for $S$ to be self-deceived in believing that $p$? I believe that the reason can be found. Mele rejects the idea that there is such a reason on the ground that self-deception should not be treated as intentional action. He holds that it is factors other than reasons, such as the desire-prompted ‘cold’ mechanism, that cause one to bias data. He thinks that there is no need to suppose that the person is aware of this in order to explain the person’s biased condition. Yet if this is the case, I want to know why $S$ biases data one way but not the other. I believe that there are appropriate and inappropriate data-biasing processes. Self-deception is arrived by biased processes that are inappropriate for the discovery of truth but which would be appropriate for attempts to conceal the truth. If this step is not taken, we will have difficulty understanding why the process is not undermined by the truth.

I believe that if $S$ is self-deceived in believing that $p$, then he believes that $p$ for a reason. It must be that $S$ believes for a wrong reason, otherwise he will be justified in believing
it, but that is an issue which need not concern us here. What does concern us is that S’s being self-deceived in believing that \( p \) cannot be satisfactorily explained by a mere ‘brute’ causal explanation. We must ask also, what reason does a rational being have for deceiving himself? What I am trying to do in Conditions 1) to 5) below is capture the reason why S deceives himself in believing that \( p \) even without having the intention to deceive himself. I think the reason explanation should be somewhat as follows:

1) S has a pro attitude towards an object \( O \),

2) S believes that the satisfaction of the object \( O \) implies that \( p \), and

3) S believes that his believing that \( p \) may bring about the satisfaction of the object \( O \), but

4) S believes that the belief that \( \neg p \) may be true, or that there is good evidence that can be counted against the belief that \( p \), and

5) S believes that, in order to justify (to himself) believing that \( p \), he must reject the belief that \( \neg p \) through concealment of the truth value relevant to \( \neg p \), or through the falsification of the belief that \( \neg p \).

Condition 1) is quite straightforward. I believe, however, S does not have to have a pro attitude towards that \( p \) to lead himself to be self-deceived into believing that \( p \). Motivated by one’s desire that \( \neg p \), one may be self-deceived into believing something that one wants the least. Clearly, one can be misled by one’s fear into believing something false but, as I argued earlier, such motivated false belief is not necessary a species of self-deception. It can be counted as a case of self-deception only it satisfy conditions 4) and 5), which I will discuss in a moment.

Condition 2) says that S believes that if he accepts that \( \neg p \) is true, then he must give up
the satisfaction of $O$. At the same time, condition 1) says that $S$ has a pro attitude towards object $O$, so that leaves no room for him to accept the belief that $\neg p$. Still, $S$ cannot believe that $p$ without doing something else first, because, Condition 4) poses an obstacle for $S$’s believing that $p$. Finally, Condition 5) says that $S$ believes that if he can convince himself that the belief that $\neg p$ is false, then he can reject the belief that $\neg p$. The moment he has rejected the belief that $\neg p$, he can justifiably believe that $p$. Consequently, he may get what he desires, namely, $O$.

The above five conditions set out to answer the two questions that a reason explanation of self-deception must deal with: First, for what reason does $S$ believe that $p$ as opposed to believing that $q$? Second, for what reason does $S$ manipulate the data the way he does?

The satisfaction of the conditions 1) through 3) gives the answer to the first question. The first three conditions constitute a sufficient reason for a conclusion, namely, $S$ intends to believe that $p$. Another way of answering the question ‘For what reason does Sid believe that Roz is playing hard to get as opposed to believe that, for instance, Roz is a straightforward kind of person?’ is this: if Sid does not believe that his desire (that Roz loves him) implies that $q$ (that Roz is a straightforward kind of person), and if $S$ does not have other reason to believe that $q$, then from Sid’s point of view, he has no reason to believe that $q$. That is, even if the belief that $q$ is true, and even if there is evidence that suggests that the belief that $q$ is true, the truth of the belief and evidence that supports the belief cannot be constituted as Sid’s reason for believing it. If Sid, for one reason or another, has not taken the evidence for the belief that $q$ into account, and if he believes that his believing that $q$ is not going to get him anywhere in his satisfaction of what he desires, then he has no reason for believing that $q$.

The relationship between Sid’s desire and his beliefs must be a logical one. Nevertheless, in the case of self-deception, the first conclusion of $S$’s reasoning, $S$’s belief that
is frustrated by condition 4), which says that $S$ has reason not to believe that $p$. $S$ needs another reason to bring about his believing that $p$. He needs a reason for acting in a certain way in order for him to believe that $p$. This reason is the reason for $S$'s manipulating data relevant to the truth value of the belief that $\neg p$. To give this reason is to give the answer to the second question, 'For what reason does $S$ manipulate data in the way he does?' The answer mainly lies in conditions 4) and 5). Note that the answer to this question is dependent on the conclusion of the first reasoning, namely, that $S$ intends to believe that $p$. Given that intention, and given condition 5), that $S$ believes that there is some way to bring the belief that $p$ about, such as by falsifying the belief that $\neg p$, together with the condition 4), we have sufficient reason for another conclusion, that is, that $S$ intentionally acts in a certain way in order to bring about his believing that $p$.

$S$ may adopt a variety of strategies to conceal or falsify the truth value relevant to the belief that $p$. They include, but are not limited to pseudo-rationalization, evasion, explaining away evidence, biassed behaviour towards the evidence against the belief held, etc. When I state the five conditions of self-deception, I give the reasons for which $S$ deceives himself. These are not usually the reasons that $S$ sees at the time he is deceiving himself. For instance, he would not see condition 5) stands the way it is. He might explain away the way that he rejects the belief that $\neg p$ as being reasonable and correct.

5. Conclusion

In this chapter, I have examined the concepts of deception and self-deception given by Mele, as well as by some other philosophers and a psychologist. I came to the conclusion that Mele's concept of deception is too broad and, hence, that Mele has over-attributed self-
deception in many cases. I have also explored the ground for deception. I believe that deception is grounded in communication. Hence my concept of deception includes the deceiver's intention to conceal or falsify the truth from the victim. Does this raise the strategy paradox again? How can one intentionally conceal truth from oneself and successfully deceive oneself in the case of self-deception? The answer is that the self-deceiver usually does not view his self-deceptive act as such. He may rationalise his act of truth concealment as a fair and insightful treatment of the data. Finally, the deceiver's concealment of truth and the formation of the self-deceptive belief need not take place simultaneously, therefore he need not simultaneously believe that $p$ and that $\neg p$.

Upon my further examination of Mele's jointly sufficient conditions of self-deception, I noticed a number of problems. I have argued that desire alone is insufficient to motivate self-deception. Desire must work in tandem with cognition in order to motivate self-deception. For instance, the agent's cognition of counter evidence must be a precondition for data manipulation. I have also discussed the concepts of desire and motivation to some extent, and how these related to self-deception.

Finally, I have pointed out the inadequacy of Mele's account of self-deception by arguing that we must give a reason-explanation as well as a causal explanation. My motive for this move is that a causal explanation of self-deception gives us no ground for holding the self-deceived person responsible, because the causal story is agentless. However, once I have argued that a self-deceiver deceives himself for a reason, then I must admit that self-deception is not, by nature, irrational. I do not see an obvious problem with this, other than to admit that I was wrong at the beginning to assume that self-deception is a form of motivated irrationality. At the end of the chapter, I have offered my own account of self-deception.
CHAPTER 4
Related Phenomena and Conclusion

1. Self-Induced Deception.

Some philosophers, such as Brian McLaughlin\(^1\) and Robert Audi\(^2\), argue that self-deception is different from self-induced deception (sometimes called self-caused deception). McLaughlin argues that self-induced deception is neither necessary nor sufficient for self-deception. First, self-induced deception is always an intentional deception, but this need not be the case in self-deception. It is necessary for wishful thinking and self-deception to be sustained by a pertinent desire, and it is necessary for both of them to be epistemically unwarranted. In contrast, self-induced deception does not require either condition to be met. That is, without desiring that \(p\), it can be the case that \(S\) is self-induced into believing that \(p\) while believing that \(p\) is warranted, although the belief that \(p\) is false. McLaughlin points out that, via memory exploitation, a person may intentionally induce himself into believing something false. Nevertheless, his believing that \(p\) is warranted. \(S\) might intentionally gather evidence which supports that \(p\) at one point, but later, when he is faced with the evidence again, he forgets why he gathered this evidence. On the basis of the evidence he has, \(S\)'s believing that \(p\) is epistemically appropriate. On the other hand, McLaughlin argues that in the case of self-deception, one’s belief is typically unwarranted.

Self-induced deception may be treated as an atypical case of intentional self-deception,

---

\(^1\)McLaughlin, 1988, pp.31-33.

as Mele did. I will argue that, aside from the warrantedness of the belief held and the intentionality of deception, one may also distinguish self-deception from self-induced deception on the basis of whether or not the agent is aware of evidence against the belief held. According to the fourth condition of my characterization of self-deception, $S$ believes that the belief that $\neg p$ may be true, or that there is good evidence that can be counted against the belief that $p$. A self-induced deceiver does not need to satisfy this condition. A self-induced deceiver need not be aware of the counter evidence and, consequently, he need not try to conceal or falsify the belief that $\neg p$ in order to induce himself to believe that $p$. Pascal's wager, for example, is such a case. Suppose that Pascal’s God does not exist. Suppose further that $S$, as the result of being persuaded by Pascal’s wager, induces himself to believe that God exists. Since the belief is arrived at by weighing the possible benefits of believing or not believing in the existence of God, the evidential support of the belief one way or the other is irrelevant to $S$’s decision. Self-induced deception, in the above sense, is more like wishful thinking in that neither requires the awareness of the fact that there is evidence against the belief held. Nevertheless, as McLaughlin rightly points out, self-induced deception is different from wishful thinking in that the latter is believing without warrantee, but the former may involve a warranted false belief. $S$ may manufacture evidence in order to lead himself to believe that $p$ in such a way that, at a later time, it seems to be epistemically appropriate for $S$ to believe that $p$ on the basis of such evidence.

Another interesting phenomenon related to self-induced deception is what I call ‘self-induced harm’, a harm one does to oneself by a self-imposed process which is (wrongly)

---

viewed as a form of self-betterment by the agent. \( S \) wants to be a certain kind of person \( X \). He believes that his doing \( A \) may change his perspective, which is necessary for his achievement of being \( X \). So he \( A \)-ed and, as the result of \( A \)-ing, he has changed his perspective. Consequently, he succeeds in being \( X \). However, if it turns out that being \( X \) is doing more harm than good to him, then what he originally took to be a plan for self-betterment is in fact a plan for self-induced harm, delusion or whatever else we name it. Worse yet, \( S \)'s new perspective brings his new ways of looking at his past, and as a result, he reinterprets the meaning of his original plan for becoming \( X \), and the nature of his \( A \)-ing also means something to him now that is different from what it meant to his in the past. As a result of such reinterpretation of his past and present, he may keep being \( X \) without even the recognition that it was he himself who induced himself into being \( X \). This, again, is not a case of self-deception, not if \( S \) does not believe, at the time of his planning to become \( X \), that the reasons against his being \( X \) are better than the reasons for it. Neither is it a case of self-deception if \( S \) neither believes nor acts upon the belief that he must conceal or falsify which might undermine his project of his becoming \( X \).

2. Wishful Thinking

Explaining wishful thinking is not nearly as complicated as explaining self-deception, and a number of philosophers seem to agree on several points with respect to the similarity between the two phenomena. Davidson,\(^4\) McLaughlin,\(^5\) Mele\(^6\) and Szabados,\(^7\) for instance, all

---

\(^4\)Davidson, 1985, pp.142-144.

\(^5\)McLaughlin, 1988, p.42.


\(^7\)Szabados, 1985, p.148.
agree that self-deception and wishful thinking are motivated by pertinent desire. Some of them agree that both self-deceivers and wishful thinkers hold a false belief, and that they are both unwarranted beliefs. Their views diverge when it comes to distinguishing the two phenomena. Davidson holds that self-deception is different from wishful thinking in that the latter needs just a wish to believe $p$ but the former requires the agent to act in certain way in order to get himself believe that $p$.\(^8\) Bela Szabados argues that the crucial difference between the two is that, in self-deception, the evidence is against the belief held, but this is not the case in wishful thinking. Secondly, the self-deceiver has good grounds for thinking that the believed proposition is false while the wishful thinker does not.\(^9\) Mele disagrees with Szabados, arguing that the wishful thinking that Szabados talks about should called ‘wishful false believing’. Mele says that the words ‘wishful thinking’ apply to more than one phenomenon. Moreover, a wishful thinker may be thinking that $p$ without actually believing it. Mele also suggests that wishful thinking may be a species of self-deception.\(^10\)

Brian McLaughlin denies that wishful thinking is a species of self-deception. He argues that, in the cases of self-deception, $S$ believes that $p$ incontinently, but in the cases of wishful thinking, $S$’s belief that $p$ is unwarranted. Moreover, typically, the self-deceiver, while believing that $p$, somewhat appreciates the evidence leads the belief that $\sim p$. This is not the case in wishful thinking. Nevertheless McLaughlin points out that, to say that wishful thinking

\(^8\)Davidson, 1985, pp.140-144.


is not a species of self-deception is not to deny that there is the continuum of cases from wishful thinking to self-deception. He believes that $S$ "moves from slight inclination to believe that not-$p$ toward actual belief that not-$p$ ... In extreme cases of self-deception, the self-deceiver simultaneously holds contradictory beliefs."\footnote{McLaughlin, 1988, p.45.}

I believe that a wishful thinker does not have to have a belief that $p$. One may be just wishing or hoping, an attitude which has much less certainty than believing, that $p$ is the case. I also do not believe that wishful thinking is a species of self-deception. Wishful thinking differs from self-deception not only in the relative strength of counter evidence, but also in that a wishful thinker, as Davidson rightly points out, does not act in a certain way to get himself to believe that $p$; he merely wishes, thinks or even believes that $p$ without trying to justify to himself why he does so. A wishful thinker must have a pro attitude towards an object $O$, he must believe that the satisfaction of the object $O$ implies that $p$, and he also must believe that his believing that $p$ may bring about the satisfaction of the object $O$, but a wishful thinker need not believe that there is good evidence that can be counted against the belief that $p$, nor need he to believe that, in order justifiably believe that $p$, he must reject the belief that $\neg p$ through concealment to, or the falsification of the truth value relevant the belief that $\neg p$ to himself. In fact, a wishful thinker is more likely than not to abandon his belief that $p$ if good evidence against that $p$ becomes available to him.

3. Naivety and Ignorance

Naivety and ignorance share some similar features with self-deception and wishful \footnotetext{McLaughlin, 1988, p.45.}
thinking. A person may have a pro attitude towards an object $O$, he may believe that the satisfaction of the object $O$ implies that $p$, and he may believe that his believing that $p$ may bring about the satisfaction of the object $O$, yet he is neither a self-deceiver nor a wishful thinker if he believes that $p$ for a reason other than the above. He may just be naive. One is naive in believing that $p$ if one does not know which direction the evidence relevant to the belief that $p$ points to. Mary, a young college student, goes to an exotic foreign country to spend her summer. Mary knows very little about the culture of that country, but she wants to be friends with the local people. She believes that, in order to be friends with them, she must be liked by them. She also believes that her believing that they like her will increase her chances of being friends with them. To her delight, she finds that the local people are very friendly to her. For instance, they smile at her whenever they see her and they often give her food and drinks. Based on this evidence, Mary believes that the local people like her. It turns out that Mary misinterpreted these people’s behaviour. They are in fact terrified by her. The smiling, the food and drinks they offer her are nothing but their way of showing her their submissiveness. Mary’s belief that the local people like her is not a case of self-deception even though she has motivation to believe this false belief. The reasons that she has, such as wanting to be friends with the local people and certain beliefs she has about how to achieve the friendship she desires, are not the ones for which she misinterprets the local people’s behaviour towards her. Mary’s misinterpretation is generated mainly by her ignorance about the meaning of these people’s behaviour. Had she understood more about the culture, the misinterpretation might have not occurred and she would not have that false belief that they like her, even though her desire to be liked by them would have remained the same.

There is a kind of false ignorance, as Annette Baier argues, that is the result of one’s
ignoring something that one knows that exists.\textsuperscript{12} When ignorance is used as a strategy for selective attention and selective evidence gathering, it is a strategy of self-deception. Mele points out that one may be ignorant about something as a result of self-deception. As Mele puts it, the failure to know the evidence may be motivated by some desire.\textsuperscript{13} It is true that a self-deceived person often does not see what is obvious; however, I would argue that not every case of ignorance is motivated or false in the way that Baier and Mele describe. In the case of Mary’s ignorance, even though she has motivations, it is not due to these motivations, but rather to the fact that she does not see where the evidence points because of her lack of background knowledge or lack of perception, that leads her to form her false belief.

This brings us to the problem of believing in the teeth of evidence, which is often described as a case of self-deception. In Mary’s case, even though from an insider’s point of view, there is abundant evidence suggesting that the local people are terrified by her, such evidence is shielded from Mary by her naivety and unmotivated ignorance. Condition 4), that $S$ believes that the belief that $\neg p$ may be true, or that there is good evidence that can be counted against the belief that $p$, is not satisfied in Mary’s case. Therefore, believing in the teeth of evidence does not suffice to constitute self-deception. Furthermore, in the teeth of evidence and with unmotivated ignorance, one may end up with a true belief. Mary may \textit{rightly} believe that a local young man, $X$, likes her because, out of ignorance, she wrongly believes that all local men like her. This is another reason for us to say that believing in the teeth of evidence does not suffice to constitute self-deception.

\textsuperscript{12}Baier, 1996, pp.53-55.

\textsuperscript{13}Mele, 1987, pp.134-5.
4. Akratic Believing and Irrationality

Akrasia, sometimes also called weakness of will or incontinence, is a name for the phenomenon of acting against one’s better judgment. Akratic believing is believing against one’s better judgment. Believing and acting are not the same sort of thing. As Mele rightly points out, our beliefs are not under direct control of our will in the same way that our actions are. Even so, beliefs are indirectly controllable by us. Not only are our beliefs based on our evidential judgments and by the choices that we make about data collection, issue focussing, etc., but what we believe can also be influenced by our evaluative judgments, as in the case of Pascal’s wager. Therefore, the explanations of akrasia and that of akratic believing can be made parallel.

There are striking similarities between akrasia and self-deception. First, both akrasia and self-deception are motivated by desire. Second, both the akratic person and the self-deceiver are conflicted. The akratic person has reasons both for and against one particular action, and the self-deceiver has evidence both for and against believing that p. Third, both akrasia and self-deception involve either unwarranted action or unwarranted believing. As Szabados rightly points out, the akratic person acts against his better judgment and the self-deceiver believes against his better evidence.¹⁴

Akrasia and self-deception are different in a number of ways. Bela Szabados argues that the akratic person is often clearly aware of the fact that he has failed to live up to his better judgment; however, the self-deceived person is blind to the fact that he is deceiving himself.¹⁵ Mele argues that self-deception is neither a necessary nor a sufficient condition for akratic


¹⁵Szabados, 1985, pp.149-151.
believing. Self-deceived belief must be false, but one may incontinently believe something true. Moreover, nonepistemic considerations in favour of holding the belief that \( p \) are irrelevant to self-deception, but are highly relevant in the case of akratic believing. First, if the belief that \( p \) is false, if \( S \)'s all-things-considered judgment is in favour of holding the belief that \( p \) and \( S \) believes that \( p \), then \( S \) is self-deceived in believing that \( p \) without believing that \( p \) incontinently. Second, if the belief that \( p \) is true, if \( S \)'s all-things-considered judgment is in favour of holding the belief that \( \neg p \), and if \( S \) believes that \( p \) anyway, then \( S \) has committed akratic believing without being self-deceived into believing that \( p \). Finally, if the belief that \( p \) is true, if \( S \) believes that the evidence supports the belief that \( \neg p \) is greater than the evidence against it, and if \( S \)'s all-things-considered judgment is in favour of holding the belief that \( p \), but \( S \) believes that \( \neg p \) anyway, then \( S \) believes that \( \neg p \) incontinently, but he is self-deceived in believing that \( \neg p \) because \( S \)'s believing that \( \neg p \) is warranted by the better evidence.\(^\text{16}\)

Let us say a few words on the irrationality of the pure cases of akrasia. Davidson points out that it is not irrational in itself if one believes that, all things considered, he ought to \( A \), but he nevertheless does \( B \). He says that all-things-considered judgments are judgments conditioned upon judging under the light of evidence that one has. There is no contradiction involved in believing against this type of judgment, if one has an unconditional judgment that tells him that he ought to believe something else. Therefore, it is not intrinsically irrational to believe against one's all-things-considered judgment. However, if the agent believes, as a principle, \( R \), that he ought to believe what his all-things-considered judgment tells him to, and if he believes that, all things considered, he ought to believe that \( p \), yet he goes on to believe

that \( \sim p \), then we have an irrational case of akratic believing. The irrationality enters when the agent is motivated to ignore acting (or believing) on his principle. However, since the agent who fails to act on his principle does not do so because he rejects the principle, there is no logical relation between his action and reason; only the causal relation between his action mental events remains. Here is one of Davidson’s famous lines: “many common examples of irrationality may be characterized by the fact that there is a mental cause that is not a reason.”

Davidson’s account of akrasia is echoed by Mele. Mele argues that an agent can believe against his decisively better epistemic judgment and thus be epistemically irresponsible without being an akratic believer, as long as he does not believe against his decisive better nonepistemic judgment. Akratic believing shows that the agent lacks self-control in the sense that, in believing the motivated belief that \( p \) during \( t \), he consciously holds a judgment to the effect that there is good and sufficient reason for his not believing that \( p \).

Now, I would argue that one’s own principle can be self-deceptively held. Suppose that the principle \( R \) (that one ought to believe what one’s all-things-considered judgment supports) is false, and suppose that a principle which is incompatible with \( R \) (say, that one ought to believe according to one’s best gut feeling) is true. Also suppose that \( S \) is motivated to believe \( R \) because he wants to be accepted by his colleagues, and believing \( R \) and acting according to \( R \) is the only way of achieving that acceptance. Suppose that \( S \) then gets himself to believe \( R \). \( S \) is self-deceived in believing that \( R \). When \( S \) cannot live up to the principle \( R \), and he believes that \( p \) (which is believing against his all-things-considered judgment), \( S \)’s akratic believing that

\( \text{17} \)Davidson, 1982, pp.294-298.

\( \text{18} \)Mele, 1987, p.112. The term ‘free*’ means that the believer is capable of controlling himself in the etiology of his holding the belief.
$p$ is, at least partly, the result of his having a self-deceived principle, $R$. Consequently, the irrationality in such cases enters when the agent is self-deceived into believing the principle $R$, which is prior to his ignoring his principle $R$ in believing that $p$. But hold on for a minute, according to Davidson, $S$ is irrational only when his belief that $R$ is held against his own standard of rationality. If so, and if $S$'s standard of rationality is so low that it lets him believe anything he wants to, then we cannot even hold him irrational for being both self-deceived into believing $R$ and incontinently believing that $p$. This seems to be problematic.

5. Lying to Oneself

Self-deception may be differentiated from lying to oneself. Lying is a form of deception, Sissela Bok defines a lie as “any intentionally deceptive message which is stated. Such statements are most often made verbally or in writing, but can of course also be conveyed via smoke signals, Morse code, sign language, and the like.”\(^{19}\) Lying may be further complicated by problems associated with verbal communication, but let us bypass these. Mele points out that the concept of lying to oneself is not problematic, but lying to oneself successfully may be hard to explain. We can understand that one believes that $p$ but tells oneself that $\sim p$ in order to achieve something else that his believing that $\sim p$ may bring about. The difficulty comes in when one succeeds in lying to oneself, as it may suggest that one believes both that $p$ and that $\sim p$. Nevertheless, Mele says that there may be, no matter how short it is, a time lag between the one’s attempt to lie to oneself and the actual state of believing that $\sim p$. By the time one believes that $\sim p$ one has already abandoned the belief that $p$.\(^{20}\) Instead

\(^{19}\)Bok, 1979, p.13.

of dividing our mind into different functional parts to deal with the problem of simultaneously believing that \( p \) and that \( \sim p \), Mele's strategy is to make a temporal division of our thought.\(^{21}\)

One successfully lies to oneself about \( p \) if only if one tells the lie with the knowledge that \( \sim p \) is warranted. This is not exactly same as in the case of intentional self-deception, but even in the case of lying to oneself, one does not have to believe both that \( p \) and that \( \sim p \), for one may have abandon the belief that \( \sim p \) when one forms the belief that \( p \). The crucial point Mele wants to make here is the that the concept of successful reflexive lying is paradoxical only if the liar believes the lie at the moment at which he lies to himself, but Mele believes that since such a concept can hardly be instantiated, we have no adequate grounds for the theoretical questioning of the concept of lying to oneself.

6. Pseudorationality \(^{22}\)

Self-deception is different from pseudorationality in that the former is arrived by virtue of the latter. Pseudorational strategies, such as denial and rationalization, are often adopted by the agent during the process of self-deception. One rationalizes or explains away one's act when one gives a reason for his action that he did not act upon.

Audi argues that

A rationalization, by \( S \), of his \( A \)-ing, is a purported account of his \( A \)-ing, given by him, which (a) offers one or more reasons for his \( A \)-ing, (b) represents his \( A \)-ing as at least

\(^{21}\)Mele realises that one cannot find any empirical ground to argue against his claim. Moreover, there seems to be a lot of room for us to explain away the problem of simultaneously holding the belief that \( p \) and that \( \sim p \), since we can attribute beliefs (such as a belief that I am a moral person) to ourselves without attributing mental activities (that is, without thinking that I am a moral person when the belief is attributed to myself, or before that). For more detailed discussion about the issue of belief attribution, see van Fraasson, 1988, pp.125-128.

\(^{22}\)Detailed discussions are made by Audi, 1993, pp.92-120 and 1988, pp.89-107; Sanford, 1988, pp.157-169; Piper, 1988, pp.297-323.
prima facie rational given the reason(s), and (c) does not explain why he A-ed.\textsuperscript{23}

When one rationalizes, one cites (or invents) reasons for an action when the cited reasons are ones that in fact play a different role in one’s action, if indeed they play any role at all. As I argued earlier, in self-deception, S has some real reasons to believe that p and to deceive himself into believing that p. The reason for the intention to believe that p, which is prior to and is included in the reason for deception, is a combination of S’s desires and beliefs: (1) S’s pro attitude towards an object O, (2) S’s belief that the satisfaction of the object O implies that p, and (3) S’s belief that his believing that p may bring about the satisfaction of the object O. The reason for S’s self-deceptive act (i.e., data-manipulation) consists, in addition to the reasons for believing that p, of two more beliefs: (1) S’s belief that ~p may be true and (2) S’s belief that, in order to justifiably believe that p, he must reject the belief that ~p through concealment of the truth value relevant to ~p, or through the falsification of the belief that ~p.

I have also argued that S need not know his real reasons for his believing that p and, in most cases, he is more likely to be unaware of them than not when he is about to be self-deceived, or is remaining in a state of being self-deceived, in believing that p. However, when asked, S may be quite readily give reasons why he believes that p. He may say that he believes that p because he believes that r implies p and that he believes that r is likely to be true. Let x be S’s belief that he believes that p because he believes that r implies p and that r is likely to be true. If S has a second-order belief that x, then B\textsuperscript{2}sx. I will argue that, although S probably does hold the second-order belief that x, B\textsuperscript{2}sx is not a real reason for which S believes that p. He is rationalizing when he cites B\textsuperscript{2}sx as his reason for believing that p because B\textsuperscript{2}sx functions

\textsuperscript{23}Audi, 1988, p.98.
differently in S's self-deception than the real reasons that I mentioned above. However, if S is
genuinely unaware of his real reasons for believing that \( p \), then his believing that \( B^2sx \), although
may contribute to his self-deceived believing that \( p \), is in itself a form of rationalization rather
than self-deception. It seems to me that a self-deceived person can hardly escape
rationalization. For the agent to know his real reasons for believing that \( p \), he would be clear-
minded about his desire and his deceptive plan, and such recognition should make him re-
evaluate his attitude toward \( p \). If so, he would be no longer believing that \( p \).

7. Conclusion

There are more questions than answers in studies of self-deception in general and in my
present work in particular. I have argued that Mele's conditions for self-deception are not
sufficient since desire alone is insufficient to motivate the agent’s data manipulation. I have
also pointed out that Mele's causal account is inadequate for explaining self-deception
satisfactorily and that reason-explanation is also required. I believe that it is important to be
able to hold the agent responsible in his self-deception. Mele's causal account, elegant as it is,
seems to me to be too 'brute' to provide a foundation for the responsibility of a self-deceived
agent. Instead, I have developed a definitive account of self-deception and of reason-
explanation. However, these are still at a rudimentary stage. No doubt there will be a lot of
room for debate as to whether the conditions of self-deception of my account are necessary or
sufficient.

I have discussed the nature of motivation and desire, and I have argued that desire and
belief must work in tandem in order to motivate the self-deceiver to manipulate data.
Obviously, my discussion of this subject has only touched the surface of the issues, and further
clarification is needed. Similarly, more study needs to be done on reason-explanation as it is relevant to self-deception. One approach may be to find a middle ground between the Humean view of acting for a reason (that is acting for achieving one’s desire) and the Kantian view (that is acting from reason).

I have discussed the concepts of deception and self-deception. I have suggested that the ground for deception lies on communication. This idea needs to be further developed. Scientific studies of communication among animals may provide some insight for our understanding of deception at its minimum level. I believe that the difference between self-deception and other-deception lies in ‘self’ and ‘other’, rather than in the concept of deception, because I believe that self-deception is a form of deception. I reject Mele’s concept of self-deception because I believe that its broadness assigns too many cases to the category of self-deception. I have pointed out one crucial feature of self-deception that Mele has missed in his account, namely that, as the result of the self-deceiver’s want and belief, he attempts to conceal or falsify the truth value relevant to the belief opposite of the one held. This feature suggests that the self-deceiver is being dishonest, not merely biased, about data relevant to the belief held. I believe that one important element of deception is dishonesty.

I started my project by assuming that self-deception is a form of motivated irrationality as Mele had suggested. Later on, I discovered that the self-deceiver who manipulates data does so for a reason that may be justified by the agent himself. I also believe that one does something irrational if one causal chain cannot be justified by the agent’s own standard of rationality, as Davidson has argued. Therefore, I came to accept that self-deception is not necessarily irrational.

My discussion of self-deception and of related phenomena such as wishful thinking and
akratic believing is not meant to be exhaustive. In particular, I think that self-deception is, instead of necessarily irrational, unavoidably pseudorational. I hope that more studies will be done on pseudorationality, where the way to a cure for self-deception may lie.
BIBLIOGRAPHY


