Fall 2010

The irrational project: toward a different understanding of self-deception

Amber Leigh Griffioen
University of Iowa

THE IRRATIONAL PROJECT: TOWARD A DIFFERENT UNDERSTANDING OF SELF-DECEPTION

by

Amber Leigh Griffioen

An Abstract

Of a thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy Degree in Philosophy in the Graduate College of The University of Iowa

December 2010

Thesis Supervisors: Associate Professor Sarah Buss
Associate Professor Evan Fales

This dissertation focuses on questions regarding the metaphysical and psychological possibility of self-deception and attempts to show that self-deception is a phenomenon best characterized as both motivated and intentional, such that self-deceivers can be held responsible for their deceptions in a stronger sense than that of being merely epistemically negligent.

In Chapter One, I introduce the paradoxes of self-deception, which arise when one attempts to draw a close analogy between self- and other-deception, and I discuss the various ways in which one might characterize an unwarranted belief as irrational. I go on to show how the various ways one understands interpersonal deception may mirror the various accounts one might give of self-deception. I conclude the chapter with a brief discussion of the role of empirical studies in philosophical investigations of irrationality.

In Chapter Two, I look more closely at a particular kind of intentionalist account of self-deception, namely the claim that we must suppose the existence of a partitioned mind to make sense of the so-called "internal irrationality" of the self-deceiver. I discuss both stronger and weaker versions of this theory, in an attempt to show that it tends to raise more metaphysical worries than it solves. I argue further that if there is such a thing as divisions within the mind, an account of self-deception centered around such divisions will not get the intentionalist about self-deception what he or she wants.

In Chapter Three, I move on to discuss non-intentionalist accounts of self-deception. Such theories have gained in popularity in recent years, due to their appeals to explanatory parsimony. Against these theories, I argued that there are certain phenomenon we take to be central to self-deception that such deflationary theories cannot account for. I therefore propose that a more robust account of self-deception is necessary to make sense of these phenomena.

Chapter Four attempts to provide such an account. I claim that if we focus more heavily on the diachronic process by which self-deceivers elicit and/or maintain their beliefs over time, what emerges looks much more like an intentional project aimed at the

manipulation of one's evidence or evidential standards than a mere more-or-less unconscious process of motivated biasing. I suggest that such a view can escape the paradoxes of self-deception, while at the same time making sense of the features lacking on non-intentionalist accounts.

Finally, in Chapter Five I examine the morality of self-deception. I argue that self-deceivers are not only epistemically but also morally responsible for their self-deceptions, and that self-deception generally represents a moral failure on the part of the moral agent, regardless of the normative moral theory one adopts.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

_____
Thesis Supervisor

_____
Title and Department

_____
Date

THE IRRATIONAL PROJECT: TOWARD A DIFFERENT UNDERSTANDING OF SELF-DECEPTION

by

Amber Leigh Griffioen

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy Degree in Philosophy in the Graduate College of The University of Iowa

December 2010

Thesis Supervisors: Associate Professor Sarah Buss
Associate Professor Evan Fales

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Amber Leigh Griffioen

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Philosophy at the December 2010 graduation.

Thesis Committee: _____
Sarah Buss, Thesis Supervisor

_____
Evan Fales, Thesis Supervisor

_____
Richard Fumerton

_____
Diane Jeske

_____
Dietrich Korsch

To *J*, who taught me to have hope without illusion.

Nothing is so difficult as not deceiving yourself.

Wittgenstein, *Culture and Value*

ACKNOWLEDGMENTS

I would first like to thank my "*Doktoreltern*": my *Doktormutter*, Dr. Sarah Buss, for introducing me to the topic of this thesis, and my *Doktorvater*, Dr. Evan Fales, for providing all the practical and philosophical assistance a *Doktorandin* could ask for. I would also like to thank my other American committee members, Dr. Richard Fumerton and Dr. Diane Jeske, for their continued support.

I must also express my gratitude to my German *Zweitgutachter*, Prof. Dr. Dietrich Korsch, for challenging me with new texts and ideas, and for pushing me to finally complete the dissertation. Your assistance was invaluable to my progress and development as a philosopher. I would also like to thank Prof. Dr. Jörg Lauster for continually raising the right objections and asking all the right questions. Without your support, this thesis would be much less interesting.

For their comments, suggestions, and help in forming my ideas on this topic, I would like to thank Neil Van Leeuwen, Maarten Boudry, Al Mele, and countless other conference attendees and referees, whose feedback has proven invaluable to my work. In this regard, I would especially like to thank Christoph Michel for his personal and professional assistance. Our many long and interesting conversations on self-deception have shaped and reshaped the way I view the phenomenon, and your friendship has been a personal buoy for me time and again.

Of course, none of this would have been possible without the continued support of family and friends. To my father, Arie, for paving the way, and to my mother, Kris, for never doubting my academic choices, I thank you from the bottom of my heart. And to all of my friends at Iowa, my deepest thanks: to Pete LeGrant for first showing me the ropes and later being my partner in crime (and for the sage advice, "Submit, submit, submit!"), to Jessica Engelking for being one of the coolest girls (and philosophers) in town, to Heather Libby for all our wonderful talks, to Kris Philipps and Seth Jones for

encouraging my hunting skills, to Shawn Akbar for being there when it counted, to Sanne and the Iowa City Germans for always playing (and living) *wilde Sau*, and to the rest of the Iowa townies, grad students, and professors, for providing so much support over the years. Finally, special thanks to my Marburger friends for helping me get by in Germany this last year: Nadine, Benny, Agnes, Jasmin, the Blues Crew, and all the others whose names I do not have room to mention here, I won't forget you.

TABLE OF CONTENTS

CHAPTER 1. EXPLORING THE POSSIBILITY OF SELF-DECEPTION: SOME PRELIMINARY DEFINITIONS AND DISTINCTIONS

1.1 Self-Deception and Interpersonal Deception: The Paradoxes of Self-Deception

The concept of 'self-deception' is a familiar one. We often speak of individuals as "deceiving", "duping", "tricking", or otherwise "lying to" themselves, as though this were no mean feat. We accuse friends, relatives, and sometimes even entire groups of people of believing something we think they ought to realize is false. Self-deception also plays a central role in works of literature, on the stage, and in film.[1] Yet when philosophers attempt to make sense of this supposedly commonplace phenomenon, they confront several puzzles. Attempting to grapple with these questions raises the question: is self-deception really possible, after all?

To a considerable extent, the difficulty of making sense of self-deception arises from the difficulty of explaining how it relates to interpersonal deception. When one person (*A*) deceives another person (*B*), *A* holds a certain proposition that *p* to be false, and he tries to bring it about that *B* comes to believe that *p*.[2] If we treat this as a model for self-deception, the agent herself must play the role of both *A* and *B*—of both the deceiv*er* and the deceiv*ed*. As the simultaneous perpetrator and victim of the deception, the self-deceived agent appears to be embroiled in paradox. Indeed, most of the philosophical literature on self-deception centers on a discussion of these paradoxes and

---

[1] See, respectively, Gustav Flaubert's *Madame Bovary* (1857), David Rabe's *Hurlyburly* (1984), and Alan Ball's *American Beauty* (1999) for a few examples.

[2] Of course, interpersonal deception may take many different forms. I discuss this possibility (and its relevance for discussions of self-deception) below. These cases (and many more) appear to fit the above description of interpersonal deception. Whether or not the more extreme cases discussed below count as instances of *deception proper* may be disputed. However, the point is merely that such cases are not difficult to make sense of, whereas their potentially self-deceptive counterparts appear—at least at first glance—to be more philosophically problematic.

how varying accounts of self-deception can (or cannot) get around them. This discussion will also be at the heart of my own account.

I will focus on two paradoxes of self-deception in particular. First, we encounter what Alfred Mele (2001) calls the "static paradox" of self-deception: if self-deception regarding *p* really is analogous to interpersonal deception, then it appears that the agent must knowingly believe a contradiction.[3] In her role as deceiver, she must believe *p* to be false, but in her role as the person deceived, she must somehow also believe *p* to be true. What's more, it appears that she must be (at least minimally) aware of this contradiction.

Second, it seems that the agent engaging in self-deception must somehow *try* to bring it about that she becomes deceived. Yet it is unclear how she can succeed if she is aware of her own deceptive intention. If, in the interpersonal case, *B* had been aware of *A*'s intention to deceive him, it is unlikely that he would have been taken in by *A*.[4] So, too, if the self-deceived agent in some sense knows what she is "up to" (in this case, trying to convince herself of the truth of *p*), it seems very unlikely that she could succeed. This has traditionally been called the "dynamic" or "strategic" paradox of self-deception.[5]

These paradoxes may lead one to conclude that self-deception cannot be analogous to interpersonal deception in any relevant sense; and this, in turn, may lead one to conclude that it is not possible for a person to deceive herself in the strong sense implied by the above analogy. Indeed, most of the differences among the philosophical accounts of self-deception reflect disagreements over whether strong self-deception is a

---

[3] Mele (2001), 7.

[4] Of course, there are cases in which we are inclined to say, e.g., that *B* "allowed himself to be taken in" by *A*. But I think that in such cases we would often be inclined to accuse *B* of some sort of *self*-deception whereby he tricks himself into believing that *A* is trustworthy.

[5] Ibid., 8.

genuine conceptual and psychological possibility. In particular, they disagree about just how closely self-deception can be said to resemble interpersonal deception, and about the extent to which the self-deceived agent is irrational. By examining these disagreements, we can gain insight into the range of phenomena that have been thought to qualify as cases of self-deception.

1.2 Delineating the Spectrum of Irrationality: The Case of Parker

An example may be useful here in helping us to properly demarcate these different kinds of cases. Let us imagine a neurologist, Parker, who is an expert in the area of Parkinson's Disease. Not only is Parker acutely able to recognize the warning signs of the disease, he is also intimately familiar with current forms of Parkinson's treatment and their respective success rates. Lately, Parker has been experiencing symptoms of a type normally indicative Parkinson's Disease. He suffers from hand tremors that temporarily disappear when he undertakes a voluntary task, he stumbles over nothing and has difficulty regaining his balance, activities requiring a degree of hand-eye coordination have become increasingly difficult, and he has begun to shuffle his feet when walking. Furthermore, he knows that Parkinson's Disease runs in his family. Call this collective body of evidence that Parker has at his disposal *E*. However, in the end, Parker fails to admit that (*q*) he may have the early stages of Parkinson's Disease. Instead, he avows that (*p*) his "symptoms" are simply caused by his being under a lot of stress at work. And suppose also that it is not the case that *p*.

A skeptic[6] regarding strong self-deception might argue that Parker is merely ignorant of or otherwise mistaken about one of the following: a) the evidence, *E*,

---

[6] I employ the word 'skeptic' here to indicate those philosophers who deny that strong self-deception is possible, such that there is *no* relevant analogy to be drawn between self-deception and interpersonal deception. This is not to say that such a philosopher would deny that the term 'self-deception' has a use in our language, but he would likely turn out to be a kind of error theorist about the phenomenon itself or would claim that, in reality, uses of the term 'self-deception' correspond to something much different than the individual parts of the term ('self'

regarding his condition itself, b) what *E* points to or implies, or c) what he himself actually believes. On this type of skeptical account, to be "self-deceived" is simply to have a false belief about oneself or one's evidence.[7] In Parker's case, it may be that he simply fails to notice *E,* or to take *E* as evidence in the first place. Or he may fail to realize that *E* implies the likelihood of *q* and thus may fail to draw the inference he ought, instead mistakenly concluding that *E* points to *p*. Finally, Parker may be mistaken about his beliefs themselves. Perhaps he publically (and perhaps even privately) avows *p*, but does not realize that he *really* believes that not-*p*—or that he believes *q* (where he also believes that *q* rules out *p* as the most plausible explanation of *E*). The point is that, on this kind of skeptical account, whenever we are inclined to attribute self-deception to an agent, we must conclude that she is either ignorant of the strength of the available evidence, or she is mistaken about what she actually believes. However, in such cases, the epistemic failure is due to a kind of unmotivated ignorance that may be fairly easily corrected.

It may also be the case that Parker cannot help but believe that *p* (or that *p* best explains *E*). Perhaps he suffers from a kind of pathological anosognosia, in which he finds himself psychologically compelled to deny that he may be suffering from Parkinson's, come what may. On this kind of skeptical account of strong self-deception, what distinguishes victims of what we call self-deception is their *resistance* to correction. This type of account claims that this resistance to correction points to the "self-deceiver"

---

and 'deception') connote. As we shall see in Chapter Three, there are philosophers who reject both of the so-called "skeptical" accounts I give below, yet who deny that self-deception is, in general, strictly analogous to interpersonal deception. (That is, they are skeptical regarding strong self-deception, but they deny that self-deception is reducible to either mere ignorance or psychological compulsion.) For these accounts, I reserve the term 'deflationary', though this term could, in principle, also apply to the skeptical theories I discuss here.

[7] There are likely other ways in which Parker may be said to be ignorant in this case, but for our purposes here, I simply wish to outline some plausible candidates.

being, at the time of her epistemic failure, under the influence of some extrinsic force (e.g. mental compulsion, disease, severe drug addiction, brainwashing, indoctrination, etc.), which prevents her from being clear about herself or her situation. According to such accounts, the self-deceived agent is compelled (psychologically or otherwise) to conclude that *p*, regardless of the strength or weakness of the evidence. Furthermore, the agent need not knowingly hold a belief to the contrary, nor need she have intentionally tried to bring about this false belief in herself. Her epistemic failure is something she is simply not competent or otherwise able to avoid.

On both of the aforementioned skeptical accounts, to say that an agent is "self-*deceived*" is, at best, misleading. This is precisely why many philosophers reject these accounts. Self-deception, they argue, is a *real* phenomenon. It involves a kind of irrationality that goes beyond mere ignorance, and it is something the agent is competent to avoid. What unites most of these philosophers is the view that self-deception is a prime example of reason "going wrong" in some way. They point out that self-deception appears to involve some failure to appreciate, evaluate, and/or employ the available evidence—a failure that disrupts the rational cognitive or epistemic processes involved in belief-formation and/or -maintenance.

Unlike skeptics about self-deception, realists stress the fact that the belief-acquisition or -maintenance of the self-deceived agent is *motivated*. According to the realist, the self-deceived agent does not just "happen" to believe irrationally. The fact that her belief is irrational has something important to do with her motivations (and, on some accounts, with her *reasons*) for believing as she does; what makes the acquisition or maintenance of certain beliefs self-deceptive has to do not only with the fact that reason goes awry in certain ways, but also that this is a response to the agent's motivations (or reasons, or both). In the case of Parker, the realist would likely argue that—if Parker really (in some sense) believes his assertions that his symptoms are not likely symptoms of Parkinson's Disease, *and* if he is not suffering from any sort of epistemic ignorance or

psycho-physical compulsion—he is guilty of a kind of *irrationality*. At the very least, he does not draw the conclusion we think he ought to draw, given the evidence he has at hand. Even if we agree on this point, however, there are multiple possible explanations of Parker's supposedly irrational behavior.

On the weaker end of the spectrum, it may be supposed that Parker is simply psychologically biased in favor of the hypothesis that his afflictions are merely symptomatic of stress, such that he does not pay attention to the evidence in ways he would were he not so biased. Two important kinds of psychological biases are frequently cited in the literature on self-deception. Parker is subject to a "cold bias" if his irrational belief is the effect of a purely cognitive mechanism, which may operate unconsciously (or more accurately, *non*-consciously) on his belief-forming processes.[8] As David Pears points out, "reason itself has certain bad habits."[9] Our tendency to pay more attention to vivid and accessible information, the propensity we have for searching out causal explanations, the confirmation bias, the gambler's fallacy, and other habitual patterns of reasoning (if we can even call it that) all seem to be "hardwired" into our brains and exert a heavy influence on many of the conclusions we draw from the available evidence.[10] Thus, "cold cases" will be those in which an agent's belief is caused in large part by the operation of one or more of these cognitive biasing mechanisms. In such cases, the agent's belief is not a rational response to the evidence, but the effect of her strong cognitive disposition to so believe.

---

[8] Note that on such an account, the mechanism in question (in this case, the survival instinct) need not actually be beneficial to the agent in all cases. Indeed, in this case it is highly unlikely that the operation of said mechanism would be beneficial to Parker, in that it leads him to avoid going to the doctor to be tested for Parkinson's and thereby to fail to receive the necessary medical treatment for his symptoms. Nevertheless, the development of such a mechanism may have been evolutionarily advantageous, even if it does not always succeed in promoting the agent's survival.

[9] Pears (1984), 9.

[10] Cf. Mele (1997), 93-4.

Alternatively, it is possible that the presence of Parker's bias is due to a (very understandable) desire that he not have the disease he has seen cripple so many of his patients over the years. In such a case of "hot biasing," Parker need not be aware of his desire and its effects on the contents of his belief. Unbeknownst to him, the desire may exert a biasing influence on his belief-forming processes, such that he comes to believe he is not ill when, absent this desire, he likely would have concluded otherwise. On the other hand, Parker may be fully aware of his desire not to be ill, and yet he may be ignorant of the fact that this desire is, to a large extent, causing him to maintain his belief that he is more or less healthy. Nevertheless, in both types of these so-called "hot cases," the agent's irrationality lies in her belief's being caused by the operations of a desire or other strong motivational state (e.g., jealousy, self-loathing, disgust, etc.)—not by a rational evaluation of the evidence she has for that belief.[11]

Philosophers who hold that most or all instances of self-deception are instances of cold or hot biasing (or both) are generally committed to the claim that this is all there is to self-deception. That is, they assert that self-deception can be explained without requiring that the agent have a relevant intention to deceive herself.[12] Thus, these so-called "non-intentionalist" accounts of self-deception maintain that although self-deception is, in fact, possible, it is not closely analogous to interpersonal deception. They claim that, unlike the case of deception between two persons, "garden-variety" cases of self-deception are

---

[11] Of course, it may also be the case that at the time of the acquisition of the belief in question, the agent *is*, in fact, rationally responsive to the evidence she takes herself to have, but that the evidence itself has been altered or distorted via the cognitive or motivational biasing mechanism prior to the agent's assessment of said evidence. I discuss this possibility further in Chapter Three.

[12] It is important to stress that, although, on this account, the agent might engage in some intentional behavior in the service of maintaining her false belief, she need not intend *to deceive herself*. Instead, the agent's belief may a product of biasing combined with certain kinds of intentional behavior. On such a view, the content of Parker's intention may not be to bring about a false belief in himself, but he does nevertheless engage in certain intentional behavior with the result that he believes he does not, in fact, have Parkinson's, due in this case to his desire not to have the disease.

not generally intentional. For this reason, proponents of such non-intentionalist views also tend to deny that self-deception involves knowingly holding contradictory beliefs. Indeed, in none of the cases described above does it appear that Parker must hold contradictory beliefs to count as self-deceived. Nevertheless, such "non-intentionalist" accounts do not usually wish to reduce self-deception to cases of sheer ignorance or psychological compulsion. Self-deceived agents, they tend to claim, have enough knowledge and competence to avoid being self-deceived. Thus, the self-deceiver may generally be said to be epistemically irresponsible in acquiring or maintaining her self-deceptive belief, despite not having a relevant intention to deceive herself.

However, some philosophers worry that those who account for self-deception in terms of hot and cold biasing avoid skepticism about self-deception at the price of applying the concept to an overly broad range of phenomena. This is, they maintain, because the accounts fail to stipulate that the agent's deception can be traced to intentional behavior of a particular kind. They overlook the fact that a self-deceiver *does* something to *ensure* that she misinterprets the evidence – that, e.g., she actively directs her attention toward or away from certain facts. According to these critics, the fact that Parker avoids discussing his symptoms with his colleagues is an essential element of his self-deception. It is not enough that he misinterpret the evidence; he must do so *willingly*.

Furthermore, one might worry that such non-intentionalist accounts cannot cover all cases of garden-variety self-deception. Indeed, one might wonder whether these accounts of self-deception really get us much further than the aforementioned skeptical accounts. Similarly, one might worry that the non-intentionalist view cannot account for certain phenomena (e.g., a certain kind of cognitive tension) that appears to arise in most epistemically irrational agents. I discuss these worries in more detail in Chapter Three.

On the other hand, accounts that attempt to preserve a strong conceptual connection between self- and other-deception tend to focus on the possibility of

intentionally acquiring or maintaining a false belief. In general, intentionalist accounts claim that either an agent can, in fact, knowingly and simultaneously hold contradictory beliefs, or that self-deception can be relevantly analogous to interpersonal deception without requiring that the agent hold contradictory beliefs. Since interpersonal deception itself may take many different forms, intentionalist accounts vary widely in other ways too. For this reason, it will be helpful here to briefly discuss some of the possible forms that interpersonal deception might take, in order to see how certain intentionalist accounts may also differ from one another.

1.3 Forms of Interpersonal Deception and Possible Analogues to Self-Deception

In simple cases, all that is needed for an individual *A* to deceive another person *B* regarding a proposition *p* is for the deceiver to merely tell a lie to the intended victim, e.g., by telling him outright that *p* is true. For example, if *B* is a young child and *A* his mother, all it may take for *A* to deceive *B* is for *A* to tell *B* that a certain false proposition is true, e.g., that the tooth fairy exists. Since the child is too inexperienced to realize that the existence of such magical entities is highly unlikely, and since his mother is one of his most trusted sources regarding the way the world is, it is likely that *B* will accept *A*'s claim about the tooth fairy without much resistance. Of course, after the initial lie, *A* may need to continue to act as though *p* is true (e.g., by putting coins under *B*'s pillow when he is sleeping), in an attempt to get *B* to continue to accept her lie.

In other cases, however, *B* might not be so easily convinced. In such cases, *A* may have to cite misleading evidence or otherwise *persuade B* to accept *p* as true. We can imagine a dishonest used car salesman who wants to convince a hesitant customer to purchase an unreliable vehicle. He may overemphasize the good qualities of the car, while glossing over certain undesirable features that would lead the customer to realize the car is not worth purchasing. In many cases, this might also involve compounding the initial untruth by telling other lies. A cheating husband may have to explain away certain

evidence that might point his wife toward his unfaithfulness or to generate a false alibi, in order to get his wife to believe that he is really faithful to her.

In more extreme cases, e.g., where *B* knows *A* to be a compulsive liar, *A* may tell *B* that *p* is *false,* hoping that *B* will then conclude *p* to be true. Alternatively, *A* may directly cause *B* to believe in the truth of *p* by, e.g., brainwashing him to believe that *p,* or by implanting a chip in his brain that would cause the belief that *p* whenever *B* entertains either the propositions *p* or its contradiction—and this, too, might well be said to count as interpersonal deception.[13]

In any of these cases, it might even be claimed that *A* deceives *B* regarding *p*, even if *p* turns out to be true: all that matters is that *A* has a deceptive *intention* to cause a false belief in *B.* For example, a deceitful lawyer may strongly believe his client to be guilty and yet attempt to bring it about that members of a jury believe his client to be innocent. And, of course, it may turn out that, unbeknownst to him, his client is actually innocent. Here, we might still accuse the lawyer of engaging in a kind of deceptive manipulation, despite the fact that, if the jury believes him, they would be believing a true proposition.[14]

Given the extreme variety of ways in which agents may be said to (relatively unproblematically) deceive each other, it is no surprise that intentionalist accounts of self-deception also differ widely, depending on what feature(s) or type(s) of interpersonal deception they take to be relevantly analogous to self-deception.

---

[13] To call such extreme cases instances of *deception* is, of course, a contentious claim. And their analogues in the realm of self-deception will be likewise controversial. (See the discussion of self-induced deception vs. self-deception below.)

[14] An even stronger example might be a case in which a policeman plants evidence to frame his rival, whom he believes to be innocent. Others may then conclude that the latter is guilty, solely on the basis of this act of deception. Of course, it may actually be the case that the policeman's rival is, in fact, guilty, independent of the evidence planted to frame him.

In the case of Parker described above, it may be that all he needs to do to deceive himself into believing that he is not exhibiting the symptoms of Parkinson's Disease is to tell a simple lie to himself and then to act in ways that support that belief. But this is easier said than done. As we have seen, the static and dynamic paradoxes appear to arise as soon as we begin to speak of the deceiver and the deceived as one person. Unlike the child lied to by his mother, Parker is highly unlikely to believe *himself*, especially if he is at the same time aware of his own deceptive intention. And even if he is somehow successful, it would seem that he must then hold contradictory beliefs, if the analogy to interpersonal deception is to hold in this case.

So, how can the intentionalist make sense of an agent's getting himself to hold contradictory beliefs in a way analogous to interpersonal deception? One suggestion has been to partition the mind into two or more subagential structures, which are capable of interacting in ways similar to distinct agents. On such a view, one or more unconscious mental structures may "act" to deceive a center of conscious agency, such that the deceptive intention is kept hidden from the agent's conscious awareness. The agent could thus be said to bring herself to believe a proposition that she also "deep down" believes to be false. Parker may have a repressed awareness of the fact that he exhibits the symptoms of Parkinson's, but this fact is somehow prevented from entering into his consciousness by an unconscious mental "censor," which, out of a (perhaps altruistic) desire to protect Parker from damaging news, deceives him into believing he is only suffering from stress. Such an account appears to preserve the analogy between interpersonal deception, while steering clear of the straightforward static and dynamic paradoxes. However, one might worry that mind-partitioning accounts solve the paradoxes of self-deception at the price of raising even more conceptual difficulties. We need to ask, for example: How do the various structures of the mind causally interact? Does an agent under these circumstances really qualify as irrational? Does she qualify as an agent at all? I return to these worries in Chapter Two.

Other intentionalists maintain that self-deception involves more than simply lying to oneself. On their view, for Parker to deceive himself, he must, in addition to intending to do so, engage in certain strategic methods of persuasion that will eventually allow him to acquire the belief he is motivated to hold. As in the case of the guileful used car salesman and the hesitant buyer, in order to persuade himself, Parker may need to rationalize or explain away certain facts he would otherwise take to be evidence against his claim that he is not exhibiting the symptoms of Parkinson's. He may need to selectively attend to certain positive evidence, e.g., that he has recently been under a lot of stress, and willfully ignore his intention to bring about what he takes to be an unwarranted belief in himself. Furthermore, as with the cheating husband above, Parker's self-deceit may give rise to a "web" of related self-deceptive beliefs, which themselves serve to maintain the initial deception. Of course, so long as Parker is actively aware of his deceptive intent, it would take a serious manipulation of both himself and the evidence he takes himself to have for him to succeed in deceiving himself. Furthermore, it appears that this manipulation must go undetected. Indeed, it appears that Parker might eventually have to forget his initial intent to deceive himself if he is ever to succeed in acquiring the false belief, unless of course he can later convince himself that his acquisition of the irrational belief is justified, despite being aware of his original intention.

Pascal's Wager may present an example of the latter kind of case. The agent who wishes to acquire a belief in God because it appears pragmatically rational (though epistemically unwarranted) may place himself in situations in which he is likely to eventually be caused to believe in God: perhaps he attends religious assemblies, reads holy texts, surrounds himself with religiously-minded people, and so on. In this way, he may come to believe that God exists for what he takes to be good reasons, despite retaining the awareness that he did not initially think it epistemically warranted to believe in the existence of a deity.

Of course, the agent who intends to deceive herself may succeed in doing so by employing more overt, memory-exploitative strategies to bring it about that at the time of the acquisition of the deceptive belief, she no longer remembers her initial intention to deceive herself. She may falsify an entry in her diary and store it for the future, knowing that 20 years later she will likely have forgotten what she really believed at the time she initially wrote the entry. Or, a patient in the early stages of Alzheimer's may, in a lucid moment, write a deceptive note to her "future" self, knowing that she will soon forget both having written the note and intending to deceive herself. On the more extreme end, someone like Parker may intentionally allow himself to be brainwashed or have his memory altered, so as to come to believe he is not ill. Such types of "self-deception" may be analogous to the more extreme cases of interpersonal deception discussed above.

However, in both kinds of cases (i.e., Pascalian cases and cases of memory exploitation), the later temporal part of the agent does not appear to be straightforwardly irrational in acquiring the supposedly false belief. At the time the Pascalian agent acquires his belief, he takes himself to have good epistemic reasons to do so, whereas before he did not. Thus, although the initial attempt may have been self-deceptive, the end result (i.e., the acquisition of the belief that God exists) is, by the agent's own lights, perfectly rational. Likewise, the diary falsifier and the Alzheimer's patient take themselves to have good reasons for their beliefs at the later time at which they acquire them. Thus, the question arises as to where the irrationality in self-deception actually lies. Moreover, is there a distinction to be made between the above types of cases? Are self-deceivers irrational in ways in which the diary falsifier and Alzheimer's patient are not? I take up the issue of what sort of irrationality is constitutive of self-deception and thus which kinds of cases count as truly self-deceptive in Chapter Four.

1.4 A Brief Outline of What Is To Come

My goal in this project is to tackle the question of what it would take for someone to deceive herself. Can we arrive at an view that both accounts for the phenomena we take to be central to self-deception and appears psychologically possible? I will examine several philosophical theories of self-deception that fall short of one or both of these objectives, insofar as they either require that we sacrifice certain of our central intuitions about self-deception and/or describe a phenomenon that seems psychologically impossible, improbable, or otherwise metaphysically suspect. I hope then to put forward an account of self-deception that meets both objectives and to show that not only is such an account conceivable, it is also likely represents a very real phenomenon.

In what follows, I will examine certain types of intentionalist and non-intentionalist accounts of self-deception, in an attempt to more fully lay out the conceptual territory regarding self-deception and self-deceptive behavior. Chapter Two will concern itself with a certain kind of intentionalist approach to self-deception, namely so-called "partitioned-mind accounts." I will examine both strong and weak accounts of this type and attempt to show that such accounts raise more worries than they solve. Further, if it turns out that these accounts do describe a real phenomenon, it will not be *intentional* self-deception—that is, the accounts will fail on their own terms.

In Chapter Three I will discuss the non-intentionalist standpoint, focusing specifically on accounts proposed by Alfred Mele and Annette Barnes. I will argue that, although the empirical literature has demonstrated the prevalence of motivational and cognitive biasing in human reasoning and belief-forming processes, such that non-intentionalist accounts of self-deception do point to very real phenomena, they fail to make sense of at least one of the possible self-relations which has a claim to being called 'self-deception'. In particular, they do not account for a stronger kind of irrationality that appears to be characteristic of self-deception and self-deceivers.

In Chapter Four, I will attempt to show that there is conceptual space for a kind of self-deceptive irrationality that is both motivated and intentional, yet does not call for a strict partitioning of the mind. After examining the techniques often employed by self-deceivers (e.g., rationalization, selective attention-directing and evidence-gathering, overcompensation, and so on), I will present the possibility of a kind of *diachronic* self-deception, viewed as a *project* or *activity* in which one intentionally engages via these means. I will argue that we have good reason to think that agents do actually exhibit this kind of irrationality, and that such a phenomenon better deserves the label 'self-deception' than do the types of phenomena examined in earlier chapters. I will further attempt to show that my account of self-deception preserves the conceptual link between self- and other-deception, while avoiding the straightforward static and dynamic paradoxes, as well as the problems that arise for views like those presented in Chapters Two and Three.

Finally, in Chapter Five, I will examine the consequences of my account of self-deception for claims about epistemic and moral responsibility. I contend that agents are, in fact, both epistemically and morally responsible for their self-deceptions. I will also argue that not only does self-deception represent a kind of epistemic failure, it also generally represents a moral failure as well, regardless of the normative ethical approach one adopts. I will conclude the chapter with suggestions for future research.

1.5 Sidenote: A Short Discussion of the Role of Empirical Studies in Philosophical Investigations of Irrationality

Before I turn to a detailed assessment of the various accounts of self-deception, a brief comment is warranted on the role that empirical studies play in philosophical treatments of self-deception. In the philosophical literature on self-deception, many references are made to studies performed by psychologists, neuroscientists, biologists,

cognitive scientists, sociologists, anthropologists, etc.[15] Scientists are understandably interested in the ways in which humans reason and form certain beliefs—especially when these processes result in the formation of beliefs we take to be in some way "irrational."[16] Indeed, the empirical literature on wishful thinking, self-deception, and other related phenomena is massive, and many scientists (and philosophers) have become increasingly suspicious of the assumption that human mental life is to a large extent rational, consistent, and under the agent's control. The studies that challenge this "rational model" of mental agency are significant and illuminating, and they provide philosophical theorists of self-deception with much fodder for discussion. However, a few words of caution are necessary here.

First, the terms 'self-deception', 'wishful thinking', 'irrational belief', 'biased belief', and so on are used quite loosely in much of the scientific (and, I would argue, philosophical) literature to indicate that the agent in question has formed a belief either that we *would* not expect her to form, given the particular scenario, or that we think she *should* not form, given the context and the evidence she has at hand.[17] As a result, one scientist's 'wishful thinking' may be another's 'self-deception'. While we do not want to reduce the debate regarding self-deception to a purely lexical disagreement, getting clear on the terms we use to delineate and describe certain psychological phenomena is, to my mind, crucial to the progress of further empirical and philosophical investigation. Thus,

---

[15] This thesis will be no exception.

[16] I put 'irrational' in scare quotes here because it is question-begging to assume that such beliefs or belief-forming processes are irrational. Indeed, it may be a purely philosophical matter as to whether a certain kind of belief, reasoning process, or action is rational or not. 'Unwarranted' or 'unjustified' may be better words to use here, though they, too, are philosophically loaded.

[17] One noticeable exception is the outstanding paper by Krizan and Windschitl (2009), in which they argue for a distinction between wishful thinking and motivated reasoning in general. They also note the limitations of empirical investigations on the desirability bias and suggest avenues for future research.

although I will try to avoid making purely stipulative claims in what follows, I will sometimes argue that certain empirical dissimilarities (especially when paired with particular conceptual-analytical considerations) may give us reason to distinguish terminologically between two phenomena or to prefer one term over another.

Second, it is important to note that empirical studies themselves do not give us the final word on what exactly is going on in the mind of the agent. In the first place, self-reports are notoriously unreliable: Agents often reinterpret or misreport their prior intentions and other psychological states—e.g., in an attempt to appear (to themselves or to others) more rational than they actually are, or to make sense of actions that appear inexplicable to them, or simply because they cannot accurately introspect or remember. Even in studies that do not rely heavily on subjects' self-reports, researchers must interpret their results and postulate causal relations that cannot be directly observed. This, quite obviously, is part of the nature of the scientific enterprise. However, it may suffice philosophically to show that a certain kind of irrationality is at least conceptually *possible*, and it is this strategy I will pursue in Chapter Four. Nevertheless, I will also argue that the phenomenological and empirical data give us good reason to think that such a phenomenon occurs regularly in the actual world.

Finally, to argue for the intelligibility and existence of irrational belief-forming processes is not to put forward the claim that we can always *know* when certain individuals have intentionally deceived themselves, as opposed to, e.g., having merely engaged in wishful thinking. Indeed, there may be no easy way of empirically distinguishing self-deceived agents from biased believers from merely ignorant agents. Similarly, the line between habituated self-deception and compulsion may be almost indistinguishable from the point of view of the third-person observer (or even from the first-person standpoint of the deceived agent herself). But this does not diminish the philosophical or pragmatic importance of making such conceptual delineations.

In this chapter, I have attempted to lay out the problems that arise when we take seriously the analogy between self-deception and interpersonal deception. I have shown the various ways in which this analogy may be drawn, and the various accounts of self-deception that have developed as a result. In the next chapter, I will discuss one very significant type of intentionalist attempt to make sense of self-deception, namely that of the aforementioned partitioned-mind theories. It is to this discussion I now turn.

CHAPTER 2. A SELF DIVIDED: PARTITIONED-MIND ACCOUNTS OF SELF-DECEPTION

2.1 Intentionalist Motivations: The Possibility of Internal Irrationality

We noted in Chapter One that there are various approaches one can take to the phenomenon of self-deception, depending on how closely one views self-deception as mirroring the deception of others. We saw that if one considers self-deception to be closely analogous to interpersonal deception, two paradoxes appear to emerge: first, the static paradox—the self-deceiver must simultaneously believe both *p* and not-*p*—and, second, the dynamic paradox— i.e., that the self-deceiver could never succeed in her deception, insofar as she is aware (*qua* deceiver) of her self-deceptive intent.

Thus it appears that any account of self-deception must adopt one of at least two strategies. It must either a) attempt to retain a close analogy with interpersonal deception that somehow escapes the static and dynamic paradoxes, or b) deny that self-deception is relevantly similar to generic other-deception, such that the aforementioned paradoxes never arise in the first place. As we have seen, non-intentionalists generally adopt the second strategy, while intentionalists tend to place significant emphasis on the 'deception' part of 'self-deception' and thus tend to adopt the first strategy. Intentionalists argue that, as in generic interpersonal deception, for an agent to count as self-*deceived*, she must, in at least some sense, intend the deception.

Tied to the intentionalist claim is the intuition that self-deception represents a paradigm case of so-called "internal irrationality." In other words, the intentionalist is often driven by the intuition that the self-deceived agent is not merely irrational from a third-person point of view but rather is irrational *by her own lights*. As Marcia Cavell puts it: "A state of mind is irrational in what is sometimes called an internal sense if it is

inconsistent or undesirable in the agent's own terms, by criteria or in light of facts he or she implicitly acknowledges."[1]

But why assume that cases of self-deception represent instances of this kind of irrationality? As Cavell further notes, "[S]omeone who typically claimed to have a belief that she acknowledged was inconsistent with her other beliefs, a belief that flew in the face of what she considered to be the available evidence, would make us doubt that she had the concept of belief; it would make us doubt her sanity, even her status as a person."[2] So why not just begin with the simpler assumption that what makes self-deception irrational is that the self-deceived agent merely fails to believe according to certain socio-cultural norms of rationality—rather than supposing that she believe against her *own* epistemic norms? Might we not do better to talk about *external* irrationality (i.e., irrationality as evaluated from an impartial third-person perspective)—for surely the self-deceiver is externally irrational, isn't she? There are several responses to this, but I will mention just a few here.

First, what makes self-deception of interest to philosophers to begin with are the two paradoxes we mentioned above. Yet both of these paradoxes are merely ways of expressing a supposed type of internal irrationality. Take a few of the many different ways of understanding the problem presented by the static paradox, for example. On the one hand, it may require that the agent straightforwardly believe a contradiction. Or it may more strongly require that she believe something she takes to be false. Or perhaps it requires that she believe something she takes herself to have no good reason to believe—or good reason to disbelieve. In any of these cases, however, it appears that the agent violates her *own* epistemic standards. They also lead to philosophically interesting

---

[1] Cavell (1998), 5.

[2] Ibid., 6.

questions regarding the dynamics of self-deception: How, if at all, is it possible to knowingly get oneself to believe a contradiction? or something one takes to be false? or something one believes oneself to have good reason not to believe? Thus, the purely conceptual question of the *possibility* of an internally irrational self-deceiver is, in and of itself, philosophically interesting. Furthermore, the answers we put forward to these questions may help us get clearer about other philosophical concepts of particular importance (e.g., the nature of beliefs and evidence, of epistemic norms, of acting for reasons, and so on).

Of course, we likely want to know more than merely whether a type of internally irrational self-deception is *conceivable*; we are also interested in discovering whether such a view fits into our conception of human action. In other words, we want our philosophical theories about self-deception to correspond to the way we understand real cases of self-deception in the actual world. And while this is in some respect an empirical matter, the task may still remain that of the philosopher to make important conceptual distinctions between various types of observed irrationality and to cash out the implications of these distinctions for such philosophical matters as the theory of mind, action theory, moral and epistemic responsibility, and so on. Thus, if we have good phenomenological, behavioral, or otherwise empirical reason to think that there is such a thing as internal irrationality, we must be able to do philosophical justice to this fact.

There is some empirical reason to think that self-deceived agents are, in fact, internally irrational, in some sense or another. One indication of this "deep" epistemic irrationality can be found in the contradictory and resistant behavior of typical self-deceivers. Self-deceived agents are not "normal" believers: they respond to evidence in ways quite uncharacteristic of typical, rational believers. When confronted with negative evidence against their beliefs (or with positive evidence for the contrary belief), self-deceivers tend to deny it more vehemently—to resist the evidence more strongly—than

do rational believers. They often become angry or defensive—sometimes attempting to rationalize their beliefs more than generally thought to be necessary.

Consider, for example, David Shapiro's observations about the behavior of self-deceived patients during therapy sessions:

> There is…[a] characteristic of [self-deceptive] speech, noticeable to the listener… When the speaker says, "I *know* I did the right thing!" or some such, with an exaggerated emphasis, one does not have a sense of being addressed. The speaker's voice is often louder than his normal conversational voice. He does not seem to be looking at one in the ordinary way. The listener does not seem to be in his focus; he seems to be looking past him. One feels tempted to wave one's hand to catch the speaker's attention. His attention seems inward, in the way of someone listening to himself, like a person who is practicing a speech.[3]

Here, we see outward evidence of what appears to be a real internal conflict within the patient. The "exaggerated emphasis" the patient places on his claim is not an indicator of certainty in what he asserts, but rather of uncertainty—of a kind of unstable tension.[4] And we might think that this cognitive tension is symptomatic of the very type of irrationality we have been interested in, namely internal irrationality. Furthermore, if internal irrationality as represented by cognitive tension in the epistemic agent really is manifested in these cases, it may give us some reason to think that the self-deceiver is not a mere victim of forces beyond his control. He is, in some sense, *complicit*, in his deception—and this may point to a level of intentionality in the phenomenon of self-deception. That is, if agents know on some level what they are up to when they are engaged in self-deception, and if this is not a result of mere psychological compulsion, then there may be a sense in which agents may be said to be *willfully* self-deceived. And this points to a kind of stronger irrationality than that of someone who merely believes

---

[3] Shapiro (1996), 789.

[4] Cf. Ibid., 792. I will discuss the notion of self-deception as an unstable condition further in Chapter Four.

against the evidence, without an awareness that they are doing so. We would not expect such agents to experience the kind of tension exhibited by Shapiro's patients, yet self-deceived agents do appear to show signs of such tension.

Finally, it is important to note that questions of internal irrationality do not only rise in examinations of self-deception. Other common forms of irrationality, e.g., incontinent action and weakness of will, also sometimes appear to exhibit a kind of internal irrationality (e.g., acting contrary to one's all-things-considered practical judgment), and if one is not a skeptic regarding the existence of these phenomena, one will have to grapple with the problem of internal irrationality here too.[5]

Thus, we may have some reason to think that internal irrationality is a real feature of self-deception—and, accordingly, we may have reason to favor the intentionalist account of self-deception. However, even if these do represent good reasons for putting forward an intentionalist view, the question remains: Can any intentionalist account of self-deception escape the above paradoxes? What might such an account look like? One popular intentionalist strategy appeals to divisions within the mind. I turn now to a discussion of this kind of theory.

2.2 An Intentionalist Strategy: Dividing the Mind

One prominent attempt employed by intentionalists to circumvent the paradoxes of self-deception is to propose a kind of mind-partitioning thesis, according to which an agent is divided into several quasi-agential, mental "substructures," at least one of which plays the role of the victimizer and another the role of the victim. Such intentionalist accounts maintain that self-deception is intentional, but only because the mind has more than one autonomous (or semi-autonomous) part.

---

[5] Of course, problems similar to those that plague intentionalist accounts of self-deception arise in explanations of these phenomena as well, especially once internal irrationality is postulated. But here, again, it is the assumption of internal irrationality that makes these phenomena philosophically interesting in the first place.

Since Freud, the notion of the divided self has made its way into ordinary parlance, such that we are often overheard to say things like, "Deep down she knows better," or even more strongly: "I subconsciously knew I loved him; I just wouldn't let myself admit it." While such a model of the mind does not immediately answer all the questions regarding the possibility of intentional self-deception, it does suggest a promising way of dealing with the static and dynamic paradoxes. That is, if we can make sense of the notion of a mind literally divided against itself, we may have a way of understanding the self as both deceiv*er* and deceiv*ed*.

2.3 Advantages of Divided-Mind Accounts

To begin, I wish to say a few words in favor of the general plausibility of such an approach. First, it is by no means a new way of thinking about the self. One need look no further than Plato's parts of the soul to see how deeply the notion of the divided self has pervaded Western philosophy. Indeed, irrationality, wrongdoing, vice, and other "shortcomings" of the individual have often been explained by an appeal to an "imbalance," "disharmony," or some other lack of unity within the self. Usually, the culprit is the so-called "appetitive" part of the soul or mind, which somehow "wins out" despite reason's best efforts to steer the agent in the right direction. Similarly, a natural way to understand the self-deceiver's apparent lack of unity is to claim that the mind itself consists of semi-independent "parts" that can fail to cohere with one another. Adopting this explanatory strategy of partitioning the mind into several (either literal or functional) quasi-autonomous structures, which themselves may serve as deceivers or victims of deceit, appears to allow the intentionalist to make sense of the self-deceiver's lack of self-coherence, without falling prey to the static or dynamic paradoxes, while preserving the analogy with other-deception. If the mind consists of several parts, it is not difficult to imagine that it might house a contradiction: the belief that *p* could be lodged in one part of the mind and the belief that not-*p* could be lodged in another.

Moreover, if these parts can interact with one another in ways similar to the interactions between agents, it appears that one part might succeed in deceiving the other in just the way that one person can deceive another.

Of course, any divided-mind account of self-deception must, minimally, answer the following questions: a) What are these various divisions, structures, or sub-systems like, and how are they demarcated? b) How do they interact with one another? c) How does such an account get rid of the paradoxes of self-deception? d) How is the agent irrational, on such an account? and, finally, e) Do we have reason to think the "self" really is divided in this way?

In the rest of this chapter, I will examine two prominent types of intentionalist partitioned-mind theories: those that employ a stronger model according to which the parts of the mind represent real structural divisions in the self, and weaker models according to which to say that there are different parts of the mind is simply to say that certain groups of mental states function as if they were discrete entities, insofar as the mental states in these groups are integrated with each other to a much greater extent than they are integrated with other mental states in the same mind. I will argue that, even if these accounts can escape the epistemic paradoxes of self-deception, they do so at the cost of raising serious metaphysical problems regarding the self, intentional action, and agency in general. Furthermore, I will attempt to show that neither model adequately makes sense of the internal irrationality the intentionalist supposes is present in self-deceived individuals. Finally, I will argue that even if we have reason to believe that mental partitioning is a real feature of (at least some) agents' mental life, this by itself is not sufficient to explain self-deception, let alone *intentional* self-deception, as the partitioned-mind theorist supposes.

I turn now to a brief discussion of Freud's model of the mind/self. This will set up my discussion of the stronger type of divided-mind theory, which I will call the "Neo-Freudian Model" of self-deception.

2.4 The Topography and Dynamics of the Mind: The Freudian Model

Although Freudian psychoanalysis has fallen out of favor with most psychologists and philosophers, Freud's theory that the mind is somehow structurally divided persists in many explanations of certain types of neuroses, "abnormal" actions, and irrationality. Of course, it seems indisputable that certain unconscious (or non-conscious) drives and mechanisms play a significant role in our mental lives, including our belief-forming processes. Many of our beliefs (or dispositions to believe) are not formed consciously and are rarely explicitly spelled out. They are often also heavily influenced by our attitudes and motivational states, many of which are not present to consciousness (as in "hot" biasing). Furthermore, cognitive (or "cold") biasing mechanisms such as confirmation and salience biases help inform the way we view the world, usually without our being aware that they are doing so.[6]

However, the intentionalist who employs a neo-Freudian model to make sense of strong cases of self-deception must go further than to merely postulate the existence of certain structural biasing mechanisms of which we are often unaware. The intentionalist must also employ the model of structural divisions in the mind to explain how it is that the self-deceiver *intends* her deception. This may, at first glance, seem unproblematic. After all, many of our commonplace explanations of self-deceptive phenomena refer to the self-deceiver's desiring, believing, or even *knowing* something unconsciously, subconsciously, or otherwise "deep down.' And quite often explanations of this type have a surprisingly intentionalist bent. We look at cases of family members who refuse to admit that incest is occurring in their own households, and we say that they must somehow *know* that something is not quite right. Indeed, we often explain this failure by claiming that such individuals *willfully ignore* the evidence of incest, perhaps by *actively*

---

[6] Cf. Nisbett & Ross (1980) for more examples of cold biases. I discuss both cold and hot biasing in more detail in the next chapter.

repressing the unwelcome evidence or belief to a place *deep down* in their psyches. Thus, even in ordinary parlance, we appear to assert that such people play an active role in their own self-deceit, with the result that they are both (subconsciously) aware and (consciously) unaware of the incest going on right under their noses.

But simply noting that we are often inclined to talk about such cases in loose Freudian terms does not provide us with good philosophical or psychological reasons to think that the mind is, in fact, so divided. We would do better to examine the Freudian model in more detail to see just how it can be employed by the neo-Freudian intentionalist to make sense of the kind of strong internal irrationality that some cases of self-deception are supposed to represent.

Prior to 1923, Freud postulated three separate "topographical" divisions of the mental sphere: the unconscious proper (*Unbewußtsein*), the preconscious (*Vorbewußtsein*), and the conscious (*Bewußtsein*). These topographical divisions were not presupposed to be dynamic and interactive, but rather merely represented different mental "provinces" or "strata", in which certain mental processes were said to occur. On this model, processes occurring in the unconscious (hereafter, *UCs*) and the preconscious (hereafter, *PCs*) are those of which the individual is not cognizant, as opposed to those of which one is directly aware, which are, quite obviously, the proper objects of consciousness (hereafter, *Cs*). Processes occurring in *PCs* were considered by Freud to be those directly "under the surface" of the conscious sphere. Thus, although *PCs* may contain objects which are not immediately present to *Cs*, they are more-or-less readily accessible to *Cs*. *UCs*, on the other hand, contains those objects (desires, beliefs, drives, impulses, etc.) which are not at all accessible to *Cs* (or at least not without significant

effort on the part of the conscious agent)—i.e., those processes which are "buried too deep" to be accessed by consciousness.[7]

After 1923, Freud added the more familiar, dynamic structures of the mind to this model, namely the id (*das Es*), the ego (*das Ich*), and the super-ego (*das Über-ich*). Of these three structures, the id was the only one supposed by Freud to be located wholly in the realm of *UCs*. Both the ego and the super-ego were considered to contain processes that occur in *Cs*, *PCs*, and *UCs*.[8] The id might be said to represent the "primordial ooze" of the mind. It contains primitive impulses and drives, and continually seeks to get its desires fulfilled. However, importantly, it does not do so via rational means. It is illogical and largely arational. It impulsively seeks the gratification of its drives at all costs. That is, it does not deliberate with reference to considerations of what is possible, given the actual circumstances of the agent, nor does it weigh competing goals, consider reasons against satisfying its drives, and so on.[9] The ego, on the other hand, represents a kind of "control apparatus."[10] Unlike the id, the ego is reality-responsive and represents a kind of "seat of rationality." It employs logical and practical reasoning, and its primary goal is self-preservation (as opposed to the id, which seeks the satisfaction of its desires at any cost).[11] Finally, the super-ego is described as the "censor" or the "conscience" of the agent. It aids in filtering out destructive impulses and is responsive to behavioral norms.

---

[7] For a more detailed description of Freud's topography of the mind, see Wolman (1968) 6-11.

[8] See Ibid., 49 for a visual diagram of this complicated later model of the mind, which also includes the "energy poles" of destructive energy and libido.

[9] Ibid., 45-8.

[10] Ibid., 50.

[11] Ibid., 54.

Thus, the super-ego acts as a sort of "sieve," through which certain impulses and thoughts are allowed for the ego to act upon, while others are "held back" or otherwise censored.[12]

We should not wonder that Freud ended up with such a complicated model of the mind. He was, at least in part, responding to the fact that agents often find themselves subject to apparently non-conscious and unintentional biases, as well as to the fact that certain "rational" processes (i.e., processes that appear responsive to certain normative principles of rationality) do not always appear to be directly present to consciousness.

However, in positing dynamic mental structures, Freud returns us to the paradoxes with which we began. If both the repressor and the mental item repressed occupy the stratum of the unconscious, then as Cavell notes, Freud "reintroduces the *unity* of deceived and deceiver now on the side of the unconscious, and it is just this unity which drives the paradoxes."[13] There is also the worry that a model of the mind that includes the ego, super-ego, and id begins to look too homuncular to be plausible. Whereas the id seems to be too disorganized and arational to count as a separate person, the "actions" purportedly perpetrated by the ego and the super-ego (e.g., repression, dissociation, deception) appear to attribute something like full-blooded agency to each of these mental substructures. But this is to make the agent-at-large look akin to a sufferer of multiple personality disorder. As David Pears notes, if a divided-mind hypothesis is the hypothesis "that there is a man within, battened down below deck like the second harpooner in Moby Dick until the moment for his action arrives,…. we might well ask how anyone could entertain such a drastic hypothesis."[14]

---

[12] Ibid., 60-5.

[13] Ibid., 8. My emphasis.

[14] Pears (1991), 397.

For these reasons, and in order to avoid unnecessary confusion, I will restrict my discussion here to *Cs*, *PCs*, and *UCs*, treating these (*pace* Freud) as both dynamic, interactive structures and topographical provinces of the mind, making little to no reference to ego, super-ego, and id. For our purposes (and, I would argue, also for Freud's), we can better make sense of intentional self-deception on this simpler, somewhat more intuitive model. I will thus move toward what is called a "depth-psychological model." One proponent of such a view, Robert Lockie (2003), writes:

> an emphasis on sexuality and a psychosexual notion of the unconscious is not essential to psychoanalysis; nor is the id/ego/superego; nor is the Oedipus Complex; and so on down a long list. … Essential to psychoanalysis [as such is merely] the notion of a *dynamic unconscious*; any position which employs such a notion is known as a *depth psychology*."[15]

I would argue that the same is true of intentionalist models of self-deception that attempt to employ a kind of Freudian divisionism. Thus, in our initial discussion of self-deception from the standpoint of partitioned-mind theory, let us employ a type of depth-psychological model on which there exist parts of the mind with quasi-independent motivational subsets that can dynamically interact with one another, at least one of which is deemed unconscious and another conscious. Both Lockie's model (2003) and Pears' very important work on motivated irrationality (1984) restrict themselves to this more limited type of neo-Freudian account, and it is these two accounts of self-deception that I will discuss in the next section.

2.5 The Neo-Freudian Model of Self-Deception: Lockie and Pears

In attempting to explain what he means by the *dynamic unconscious*, Lockie (2003) makes the following claims:

- The person is made up of *parts* of some kind.

---

[15] Lockie (2003), 127-8.

- The parts have their own motivational interests (desires, bio-psychological drives, socio-psychological norms and ideals).
- The activities and motivations of these parts are not necessarily known to, or shared by, the other parts, or the "person as a whole."[16]
- The sense in which these parts may have or lack knowledge of the other parts, or the person as a whole is an *active* sense. So, these parts have (partially successful) means of selectively concealing, revealing, and deceiving the other parts, and in other ways competing for psychological resources, in accordance with their (the parts') sub-interests. This internal psychical "ecology" is an *active system* of interacting parts (involving processes of competition, cooperation, symbiosis, etc.); it is in this sense that it is "dynamic."[17]

Given Pear's discussion of self-deception in his 1984 book, *Motivated Irrationality*, I take it that he would not be opposed to characterizing the "sub-intentional structures" of the mind in such a way. Thus, on the neo-Freudian depth-psychological model under discussion, when we speak of "parts" of the mind, it is important to remember that something like the above description is what we have in mind.[18]

Pears claims that most cases of "hot" motivational biasing, as well as the potentially stronger cases of belief-manipulation we mentioned in Chapter One, support the hypothesis that the agent does not have a single, unified consciousness. For example, in order for a persona's wish to influence her belief, she cannot be consciously aware of this influence:

> A person's beliefs adjust themselves directly to his evidence and a wish cannot simply stand between his evidence and his beliefs like a policeman directing traffic. …. If we are going to identify and ascribe a strategy in cases of this kind, we shall have to ascribe it to a sub-system within the person. Such a sub-system would

---

[16] A better substitution for "the person as a whole" might be "the conscious system," at least on a Pearsian account of the mind.

[17] Ibid., 128. I have intentionally omitted Lockie's final criterion, as it is one with which I am not sure Pears would agree. It reads: "The activities of these parts may often be revealed only obliquely, by skilled inference (psychological "detective work") to uncover the meaning revealed in slips, dreams, bungled actions, selective amnesia, neurotic symptoms, character traits, etc. (typically by using clinical interview techniques—free association, hypnosis, and other techniques allied to the clinical experience)."

[18] No pun intended.

> confront the rest of the person in something like the way in which the whole person confronts another person. It might even notice the weaknesses in the rest of the person and devise strategies to exploit them.[19]

In other words, Pears claims that there must be mental divisions of the type described by Lockie, which themselves are instrumental in something's being (in at least some cases, intentionally) kept from the agent's consciousness, thereby allowing her to acquire or maintain a certain belief.

What, exactly, does the self-deceived individual look like on such a model? According to the self-acknowledged "Freudian principle" to which Pears appeals, "the main system include[s] everything accessible to a person's consciousness, while the sub-system[s] include[s] everything else that is needed to explain his speech and behaviour."[20] On this view, if we cannot locate the cause of the formation and/or maintenance of an irrational belief in an agent's consciousness (i.e., the main system), there must be at least one sub-system, the existence of which can explain the agent's holding that belief. Whereas Freud is primarily concerned with deeply repressed *wishes* located in the largely chaotic region of *UCs*, Pears points out that self-deception often requires only a kind of "shallow" repression of a certain belief into *PCs*, and "even that only needs to last as long as is necessary for the particular piece of self-deception."[21] Additionally, as we have seen, the activities of *UCs* on the Freudian model have little to no rational structure and are more or less indifferent to any restrictions imposed by the agent's conception of reality, but self-deceived agents are often engaged in very complicated means-ends strategies that do appear to reflect a certain degree of practical rationality and reality-responsiveness. Thus, many instances of self-deception on the

---

[19] Pears (1984), 63.

[20] Ibid. 68.

[21] Ibid. 74.

Pearsian model will merely involve the interaction of the *PCs* and *Cs* systems, without the aid of *UCs*.[22]

Pears notes that, for Freud, the *permissive cause* of any particular piece of epistemic irrationality—where by 'permissive cause,' Pears appears to mean something like a kind of "passivity" on the part of some mental substructure—is always represented as a failure of consciousness, e.g., the failure to acknowledge something, the realization of which would likely lead to the agent's believing a different proposition. For Pears, this "something" is a kind of cautionary belief, that is, a second-order belief about the illegitimacy of the formation or maintenance of the irrational belief.[23] For example, the permissive cause of an agent's believing irrationally that *p* may be her failure to become or remain consciously aware of the cautionary belief that she possesses ample evidence against *p*, or that her desire that *p* is causing the belief that *p*. (Presumably, if the agent were consciously aware of either of these beliefs, she would revise her belief and conclude that not-*p*.[24])

---

[22] Of course, some cases of strong biasing may be the product of a desire or other impulse in *UCs* exerting its strength in *Cs*. But this does not seem "intentional," in the sense that the intentionalist partitioned-mind theorist is looking for. As mentioned, if *UCs* is incapable of rational, intentional behavior, it does not seem that the origin of intentional self-deception could be located in that region.

[23] One might think that what must be kept from *Cs* is the presence of the motivating factor itself, which is illicitly causing the irrational belief. However, this is often not the case with self-deceived individuals. Indeed, the agent motivated by a desire to believe that *p* may be perfectly aware of her desire that *p* and even of her desire to believe that *p*. Rather, what must be "kept" out of *Cs* is, on the Pearsian model, the belief that this desire is actually causing her belief that *p*.

[24] One might object here that in the case where one becomes conscious of one's belief that the evidence points to not-*p*'s being the case, one might still intentionally fail to conclude that not-*p*. Indeed, there is always some "latitude" in inductive reasoning (cf. Pears, 1984, 76-7 and Davidson, 1980), since there is always an "epistemic gap" on the part of the doxastic agent between the conclusion she is inclined to draw from the evidence she has and the conclusion that she would draw if one had *all* the evidence. Thus, we can imagine an agent refusing to accept a certain conclusion, concluding instead that "not all the evidence has been considered." However, if the agent does so irrationally, it would seem that, on the Pearsian model being considered here, *Cs* would still fail to acknowledge something—in this case, perhaps, the belief that a desire to

Such a model points out something very important about what intentional self-deception must look like—namely, the self-deceived agent does not merely treat evidence in a biased fashion. She also has a *second-order* belief about her beliefs or motivations. In short, if self-deception really is to be understood on an intentional model, then it appears the agent must have some *reflective* understanding of the fact that her belief is being deviantly caused. The mere fact that an agent's, e.g., first-order desire (as opposed to a sensitivity to the evidence for or against *p*'s being true) is causing her belief that *p* is not sufficient for the agent's counting as self-deceived. Rather, she must, in some sense, "know" what she is up to. And it is this reflective acknowledgement that must be relegated to the non-conscious sub-system (e.g., to *PCs*), in order for the particular piece of self-deception to be successful.[25]

Of course, as Pears points out, the failure of consciousness to acknowledge a cautionary belief is also not sufficient to explain self-deceptive irrationality. The depth-psychological model must also explain *why* this preconscious cautionary belief does not produce the same effects on *Cs* when housed in *PCs* as it does when it is in *Cs*. In *Cs*, the cautionary belief would likely motivate the agent to revise her belief. However, once it is relegated to *PCs*, it no longer appears able to do so. Pears puts the problem this way:

> [I]f the lapse of the cautionary belief from consciousness did provide a complete explanation of its non-intervention, that would be because preconscious beliefs are always powerless to produce their normal effects in consciousness. …[A] preconscious belief would always be shut up in its sub-system and it would never be able to use the sub-system as a base for operations in the main system. But this is not universally true [of] preconscious beliefs.[26]

---

believe that *p* (as opposed to an inclination to engage in "responsible" inductive reasoning) is motivating this refusal to believe that not-*p*.

[25] I will take up this point in some detail later on.

[26] Pears, op. cit., 79.

On the Freudian model, mental states in *PCs* very often affect one's actions. Though we naturally filter out certain "distracting" input, e.g., in order to focus on a particular matter at hand, we are often responsive to such filtered stimuli—that is, they affect our conscious activity. Even in cases of self-deception, however, the cautionary belief does not lose all power over the agent's conscious activity. Self-deceived individuals appear to engage in attention directing, active avoidance of certain evidence, and other responses to cognitive conflict. So the cautionary belief must be doing *something* from its "base" in *PCs*. How can we explain this?

Pears claims that placing all our emphasis on the permissive cause of irrationality will not allow us to explain how it is that individuals may become self-deceived. Rather, we must further inquire as to the *productive cause(s)* of self-deception on such a model. That is, we need an account of the mental state that serves, as Davidson puts it, as a cause that is not a reason for what it causes—not just an account of the *thing* that "allows" the irrationality to occur.[27] For Freud, the *productive cause* of self-deception and other types of irrational belief formation is always centered around a wish.[28] However, Pears notes that there are other possibilities. "Perversions" or "bad habits" of reason, such as confirmation and salience biases, may also lead to failures of consciousness to appreciate what it ought.[29] Additionally, it seems that emotions other than wishes, such as feelings of extreme fear or self-loathing may also play significant productive causal roles in specific types of self-deception.[30] Thus, what emerges is a modification of Freud's

---

[27] Cf. Gardner (1993), 65.

[28] Pears, op. cit., 71.

[29] Ibid. (Cf. also p.9.)

[30] Here one might think of, e.g., a perpetually jealous husband, who constantly deceives himself that his wife is unfaithful to him, despite quite obviously not wishing this to be the case. The motivation in such a case may be something like self-loathing or a fear that his wife "settled" when she married him, etc.

theory, according to which epistemic irrationality is the result of a *permissive* failure of conscious awareness, where this failure is largely *produced* by some sort of biasing mechanism, desire, or strong emotion. And Pears supposes that such a thing could not occur without postulating separate systems in the mind, which interact in ways similar to two people.

A shift to a discussion of the productive cause(s) of self-deception, combined with a commitment to the quasi-intersubjective nature of mental systems, may also aid the Pearsian intentionalist in explaining how it is that the cautionary belief cannot produce its normal effects in *Cs* once it has been relegated to *PCs*. If the cautionary belief is centered around some motivational element (say, the desire that *Cs* believe that *p*)—and if *PCs* interacts with *Cs* in a way similar to how two people interact—then it is perfectly conceivable that it is the presence of this desire in *PCs* that serves as the nucleus of the activity (or lack thereof) of the cautionary belief in *PCs* on *Cs*. It is in this way that *PCs* may undertake a kind of intentional deception of *Cs*, whereby the former keeps certain information (in this case, the second-order, cautionary belief that the conscious belief that *p* was formed irrationally) from the latter, due to the existence of a certain (perhaps altruistic) desire in *PCs* to do so.

Additionally, the fact that the cautionary belief and the deceptive desire are housed in *PCs*, as opposed to *UCs*, may explain the cognitive tension the self-deceived individual often experiences or exhibits in cases of strong self-deception (where the productive cause of the deception is not a mere motivational or cognitive biasing mechanism). Since *PCs* contains information that is "just beneath the surface" of consciousness, so to speak, the neo-Freudian can make sense of the fact that the cautionary belief and/or *PCs*'s desire to deceive may occasionally surface at the level of *Cs*. Since the information in *PCs* is often readily available to consciousness, if an agent's conscious awareness were turned in that direction, one might expect that her behavior would sometimes be "symptomatic" of inner conflict. For example, the emergence of a

new piece of evidence against her self-deceptive belief might briefly cause her to turn her attention to the cautionary belief in *PCs* (thereby "transporting" the cautionary belief into or "sharing" it with *Cs*), and she might then exhibit signs of anxiety or doubt with respect to her irrational belief. However, if *PCs* is somehow capable of "re-hiding" the cautionary belief from *Cs*, then it is possible she will be able to maintain the deception, having only exhibited a momentary symptom of, or lapse in, her self-deception. As Pears writes in a later article:

> It is not necessary for a self-deceiver to be [perpetually] unaware of the existence of the elements that go to form the sub-system within him. All that is necessary is that he should not realise or avow the self-deceptive intention at the time that it is being executed.[31]

We can think about this phenomenon more clearly if we compare it to instances of interpersonal deception. In cases of the sustained deception of one person by another—for example, in a situation in which a con artist is attempting to swindle his "mark" out of a large sum of money—it often happens that the victim of the deception becomes suspicious of the perpetrator of the deception. In cases of the so-called "long con," the intended victim may hear from a friend that the swindler is untrustworthy. This may cause the victim to question her faith in the swindler. She may notice things about him that she had not previously noticed—a shifty glance, strange behavior, commiseration with undesirable individuals, and so on. Thus, her behavior toward the swindler may change noticeably. However, if the swindler is a skillful con artist, he will have the guile and deceptive wherewithal to reestablish the trust of his mark, and will thereby be able to maintain the deception. Presumably, however, the maintenance of the deception will also rely on some complicity on the part of the intended victim. Thus, we can compare this

---

[31] Pears (1991) 399-400.

continued effort by the swindler to deceive the mark to the actions of *PCs* toward *Cs* and the complicity of the mark toward the swindler to the permissiveness of *Cs* toward *PCs*.

To tie all these components of Pears' neo-Freudian model together, let us see how the model might be applied to a case in which an agent appears to intentionally deceive herself. Agnes loves her husband, Ralph, more than anything else. That is, his love is extremely important to her, and his faithfulness to her is a reflection of this love. In the past, she has had no reason to question his fidelity, but over the past couple of months his behavior has become erratic. He comes home late at night smelling of cheap perfume, making implausible excuses for his tardiness; she has found lipstick stains on his collar; he has shown little to no sexual interest in her; etc. Suppose, too, that Agnes' friends have informed her that they have seen Ralph dining with another woman at a romantic restaurant in the area. At this point, Agnes has more than enough evidence to suspect that Ralph is being unfaithful, and indeed, her own assessment of the evidence points toward his having an affair. However, as any loving wife would, Agnes also has a strong desire that he *not* be cheating on her. So let us suppose that, under the influence of this desire, she refuses to admit that he is unfaithful, even to herself. In other words, she self-deceptively attempts to maintain her belief that he is faithful, despite the fact that her evidence suggests otherwise. Suppose also that she exhibits signs of significant cognitive tension, suggesting that she is, in fact, "at odds" with herself in some significant way. She refuses to listen to her friends when they confront her about Ralph, and continually tries to change the subject. Yet when forced to listen, she persistently defends Ralph and asserts in an overly strong manner that he is faithful. She avoids social situations that might lead her to encounter Ralph in a compromising position, and she invents excuses for his strange behavior. This may lead us to believe that that she really is aware *on some level* that her maintenance of this belief is irrational. But how can we explain her behavior without running into the paradoxes of self-deception?

As we have seen, if this case is to amount to something more than mere wishful thinking, in which Agnes' desire that Ralph be faithful causally sustains her belief that he is not cheating on her (where she cannot be said to know—on any level—about the operation of her belief-sustaining desire), we must postulate that she is somehow both aware and unaware of what she is doing. It is here that the depth-psychological model comes into play. Pears would suggest that although Agnes knows that she possesses good evidence that her belief in Ralph's fidelity is likely false, and that it is merely her desire that he be faithful that causally sustains her belief in his fidelity, this fact is actively kept out of her conscious awareness by the functioning of her *PCs* system, which—being built around her wish that Ralph be faithful (and, perhaps also around the wish on the part of her *PCs* that she not be consciously aware of this fact)—hides the potentially traumatic information from her *Cs* system. Thus, Agnes might be said to both (preconsciously) know and (consciously) not to know that her belief in Ralph's fidelity is being irrationally sustained. In such a case, it does not appear that Agnes consciously believes both that *p* and that not-*p* (where *p* represents the belief that Ralph is faithful and not-*p* represents the belief that he is not faithful). Indeed, it need not be the case that her *PCs* system ever forms the belief that Ralph is unfaithful (or that it is not the case that he is faithful). It merely prevents her *Cs* system from becoming aware of certain information—in this case, of the cautionary belief that her belief is being irrationally sustained. In other words, *PCs* withholds certain information from *Cs*, which—if she were consciously aware of it—would normally cause her to revise her belief. The partitioned-mind model thus avoids the straightforward static paradox. Since Agnes' belief that *p* belongs to a separate system than the cautionary belief, the presence in her of inconsistent beliefs appears no more paradoxical than the presence of such beliefs in two separate individuals.

It is less clear that the dynamic depth-psychological model has the resources to avoid the dynamic paradox. At first glance, it appears to have a simple answer to the question of how Agnes can be successful in her self-deception if she is aware of her own

deceptive intention—namely, that she is *not* directly aware of her deceptive intention at the time she is deceived. The deceptive intention is kept out of her conscious system long enough to do its self-deceptive damage. Thus, just as with interpersonal deception, there is no paradox.

But the problem is not so easily side-stepped. Although Pears has told us that the *UCs*, *PCs*, and *Cs* systems may work together to keep a particular belief or set of beliefs out of *Cs*, a more detailed discussion of exactly *how* these systems interact will be necessary for the depth-psychological model to give us an account of self-deception a) that is, in some *relevant* and *plausible* sense, intentional, and b) that preserves the sense of internal irrationality discussed above.

First, there is the question of what exactly must be kept from consciousness, in order for the deception to succeed. At the very least, the cautionary belief that the self-deceptive belief is primarily caused by a desire or some other element in the agent's motivational set must be prevented from entering *Cs*, in order to prevent *Cs* from revising its self-deceptive belief. But in some cases, the motivation itself may need to be repressed. For example, if the motivation behind an agent's particular piece of self-deception is not the desire that things be a certain way, but rather the desire *to believe* a certain proposition (perhaps because she takes the belief in that proposition to have certain practical benefits), knowing that she desires to believe it might count as evidence against this belief, especially if she is aware of additional evidence pointing in the same direction. Thus, the self-deception in question might never get off the ground, if the agent is aware of her motivation for so believing.[32] The awareness of other related motivations

---

[32] This might be especially true of destructive motivations, such as self-hatred. The mere awareness of the fact that one *has* such a motivation might lead the conscious system to suspect that this motivation plays a large factor in her believing a particular proposition (e.g., that a certain person dislikes her)—where such an awareness would precede (and indeed lead to) an awareness of the "cautionary belief" that the motivation is, in fact, exercising this causal power on her belief system.

may need to be repressed as well. Take, for example, an agent who has homosexual impulses stemming from *UCs* but who desires not to be a homosexual. While awareness of the latter desire (which is doing much of the causal work in the formation of the self-deceptive belief) need not be kept from *Cs*, the former likely must—if the deception is to succeed.[33]

Then there is the question as to the relevant intentions involved in intentional self-deception. Presumably, any intention to deceive oneself must also be kept from *Cs*, in order to allow the deception to occur. This leads to a related question regarding which intention or intentions are the relevant one(s), as far as self-deception is concerned. The obvious answer to this question would be that it is the agent's intention to deceive herself (or to believe something unwarranted by the evidence), where this intention is (at least temporarily) prevented from entering *Cs* by *PCs*. But if this intention is confined to *PCs*, then how can it be the *agent's* intention (if we understand the agent here as being represented by *Cs*)? Assuming as we did that the unruly *UCs* (if it exists at all) is incapable of anything like intention, it appears deceptive intention must be on the part of the *PCs* sub-system and directed at the main system. This appears consistent with what Pears says when he explains why self-deceivers typically cannot produce descriptions of their belief-forming processes from the agent's (or *Cs*'s) point of view.[34] He writes: "This, of course, is no surprise, because it is not [the agent's] basic act but the sub-

---

[33] Furthermore, as we shall see shortly, *PCs* must have its own motivations for keeping certain beliefs from *Cs*, and presumably these motivations must themselves be kept from *Cs*.

[34] I am not sure, however, that this is quite right. Self-deceivers can, and often do, produce rationalizing descriptions of their behavior. However, what I take Pears to mean is that self-deceivers cannot produce *accurate* descriptions of how they came to form their self-deceptive belief. There is some sense in which the self-deceiver cannot understand himself. And, presumably, it is this sense in which the agent is supposed to be internally irrational. But even if this lack of self-understanding explains some of the tension displayed by self-deceivers, I am not sure it is enough to get us deep epistemic irrationality, since, as far as *Cs* is concerned, it believes for what it takes to be good reasons. I elaborate on this point below.

system's, and the sub-system does do it under a description, namely, the description 'generating the counter-evidential belief preferred by the main system'."[35]

Pears here refers to two potential motivations for the agent's forming or sustaining a self-deceptive belief. First, there is the motivation in the form of a preference in the *main system* (or, for our purposes, in *Cs*). And this preference need not be withheld from *Cs*, though in some cases it may be (as we have seen above). Second, there is the motivation of the sub-system (or *PCs*). If *PCs* itself is capable of intentionally "generating the counter-evidential belief preferred by the main system," it must do so for a *reason*—that is, it must have its own motivations for strategically generating this belief in *Cs*. But what might those motivations be? Are they altruistic desires—designed to protect the agent from herself? to preserve her self-concept? to reduce anxiety? Surely there are cases in which it makes little sense to suppose that *PCs*'s motivation is of this kind.[36] Or does *PCs* act according to a sort of Freudian "pleasure principle," which seeks pleasure for *Cs* at all costs, rather than deferring gratification when necessary? But this would make PCs's intention appear significantly less rational than we initially took it to be.

Since Pears describes the intentional self-deceptive *act* as belonging to the *sub-system* in question, or *PCs* on our depth-psychological model, it would seem that *PCs* must possess at least enough agency[37] to be able to form and execute intentions, such as the intention to deceive *Cs*. But if *PCs* really can be said to intentionally cause or

---

[35] Ibid. 401.

[36] Take, for example, the partygoer who self-deceptively convinces himself that it is all right to have one more drink before driving home. In a case such as this, it is rather implausible to think that *PCs*'s motivation is aimed at the agent's well-being.

[37] Pears discusses agency as a gradualistic concept, as opposed to taking an all-or-nothing approach. I am actually sympathetic to this idea, though as I will discuss shortly, I do not think it cannot save Pears from certain homuncularist worries.

maintain a false belief in *Cs*, then it must have a significant awareness of what is going on in *Cs* and what it must do to protect *Cs* from coming to an awareness of the relevant beliefs, motivations, and intentions in question. Indeed, it seems that in order to make sense of how *PCs* can do so, we must attribute to it a rather strong sense of agency, and this makes Pears' account appear more homuncular than we initially proposed. Yet Pears denies this:

> The picture of the tiny agent within must be dismissed, because this kind of systemic explanation is not homuncular but anthropoid. It does not credit the sub-system with *all* the features of ordinary agency on a *reduced* scale, but only with *some* of them, and those on the *full* scale. …. What is being suggested is that the conscious processing of information may be paralleled by an equally effective kind of processing which does not require a separate centre of consciousness, but which is, nevertheless, sufficient to support the concept of a separate centre of agency[38]

However, Pears' protestations to the contrary notwithstanding, it is rather difficult to see how the notion of a "separate centre of agency" is even coherent, unless the sub-system can be said to really resemble conscious agency. He suggests the existence of a kind of processing analogous to conscious reasoning that occurs on a pre-conscious (and thus *non*-conscious) level but which can produce results in *Cs*. I do not wish to claim that there is can be no such thing as non-conscious cognitive processing that influences one's conscious beliefs or actions. As we have seen, the prevalence of such phenomena as hot and cold biasing in normal agents points to the idea that there may be unconscious mental states that can produce effects in consciousness. What I wish to question is why we should suppose that this kind of non-conscious cognitive processing allows for the attribution of any kind of intentional, rational agency to the sub-system in question.

If intentional action is, in some sense, action for a reason, then it appears that *PCs* must also be able to appreciate, evaluate, and act for reasons. But this appears to require a

---

[38] Ibid. 398, 400.

level of (potential) reflectivity and awareness on the part of *PCs*. And Pears does appear to hold that the sub-system has some sort of self-knowledge: "We use behavioural evidence for the conclusion that the sub-system knows what it is doing and why it is doing it, and is aware of its own success, if it is succeeding."[39] Yet Pears also claims that precisely what is lacking in the sub-system is a type of conscious awareness: "What is lacking in the case of a self-deceptive sub-system is not only the contemporary avowal of its intention but also a detailed description of its basic actions."[40] Thus Pears appears to maintain that we can attribute an intention (and thus agency) to *PCs* because it "knows what it is doing and why it is doing it," but at the same time *PCs* does not have the requisite awareness to avow its own intention or describe what it is doing. This seems problematic. How can *PCs* reflectively know what it is doing if it is incapable of that level of reflection?

One possibility is that *PCs* itself acts for reasons that are not available to it. In fact, its "reasons" cannot be conscious, given that *PCs* itself lacks this important feature of agency. But then we appear to be left with an unpleasant alternative. In order to make sense of why *PCs* does what it does, we may need to appeal to further centers of agency, which themselves might house *PCs* "reasons" for deception, to explain the "actions" of *PCs* itself. But then we raise the threat of a regress, and even if this regress is not vicious, at the very least we must worry about losing the agent altogether. As Raziel Abelson writes:

> The agent appears as a passive victim of competing forces within his unconscious and is no longer responsible for his actions. But then they are not *his* actions, and the agent has been reduced to a mere battlefield—he is no longer an agent.[41]

---

[39] Ibid. 400.

[40] Ibid. 403.

[41] Abelson (1977), 97.

Thus, if we try to rescue Pears from falling into a sort of implausible homuncularist view, we do not appear to be able to make sense of how self-deception is intentional in the first place because *PCs* begins to look like a mere causal mechanism altogether incapable of intentional action (or at least like the non-conscious product of such mechanisms). And this would be to characterize self-deception as significantly less intentional than Pears would have it.

In the end, Pears himself admits: "We have a general description of what [the sub-system] does: it exerts a continuous influence on the main system's thinking. But we do not know how it does this because it is a task too far removed from the ordinary things which we do intentionally, and which we can break down into their basic elements."[42] He goes on to claim that this is no reason to conclude that the sub-system does not represent a separate center of agency, but if understanding how *PCs* operates requires diverging too far from comprehensible intentional behavior, then it does not appear to warrant any attribution of agency to it.

Furthermore, Pear's view ends up threatening the requirement of internal irrationality for self-deception that motivated the account in the first place. If the "agent" is to be identified with *Cs*, then her self-deceptive belief is merely the effect of whichever system or systems causally win out in the "belief battle." But the conscious agent takes herself to believe for good epistemic reasons (as provided by *PCs*). Moreover, it is difficult to see why the agent would experience any cognitive tension in believing as she does, insofar as she is not aware of the inconsistent beliefs housed elsewhere in her (unconscious) mind.

For these reasons, it may be reasonable to look for a weaker divisionist thesis than that proposed by the neo-Freudian depth-psychological model—one which attempts to

---

[42] Pears, op. cit.

explain internal irrationality without resorting to homuncularism or purely mechanistic views. It is to such a view that I now turn.

2.6 Davidson's "Functional" Model.

In his seminal 1982 article "Paradoxes of Irrationality," Donald Davidson rejects the neo-Freudian divisionist accounts, while nonetheless arguing that we need to suppose a partitioned mind if we are to do justice to the irrationality of self-deception. Like Pears, Davidson is motivated to put forward an account on which self-deception is *internally* irrational. In "Incoherence and Irrationality," he writes:

> No doubt we very often stigmatize an action, belief, attitude, or piece of reasoning as irrational simply because we disapprove, disagree, are offended, or find something not up to our own standards. … [However,] my interest here is entirely with cases, if such there be, in which the judgment that the works or thoughts of an agent are irrational is not based, or at least not necessarily based, on disagreement over fact or norm. … This suggests that we should limit ourselves to cases in which an agent acts, thinks, or feels counter to his own conception of what is reasonable; cases where there is some sort of inner inconsistency or incoherence.[43]

Furthermore, this interest in cases of internal irrationality prompts Davidson to focus on cases in which self-deception can be said to be *intentional*, for, he thinks, if self-deception can be chalked up to mere ignorance or to some sort of mental compulsion, then the agent is in no way criticizably irrational, even if she is not responsive to the reasons she has.

Finally, this concern to explicate cases of intentional, internal irrationality motivates Davidson to distinguish between cases of irrationality motivated purely by a desire to so believe (i.e., wishful thinking) and self-deception proper: "When wishful thinking…succeeds there is no moment at which the thinker must be irrational. …. Not all wishful thinking is self-deception, since the latter but not the former requires

---

[43] In Davidson (2004), 189.

intervention by the agent."[44] Thus, Davidson is interested in giving an account of self-deception that allows for the internal irrationality of the agent in question, which presupposes that the agent deceives herself intentionally, yet nevertheless does not reduce self-deception to a species of wishful thinking.[45] What does such an account look like?

We begin with the claim that agents believe for reasons. And one's reasons bear certain logical and causal relations to the formation of the belief in question. Thus, when the self-deceiver forms her irrational belief, although she does so for reasons, she believes contrary to those reasons she takes to be the strongest. Take Agnes, for example. She has reasons both for and against her belief that Ralph is faithful to her. However, in believing that he is faithful, she violates a principle she otherwise accepts: the "requirement of total evidence", i.e., the requirement that one ought to believe on the totality of one's evidence. She believes a proposition for whose contrary she takes herself to have *better* evidence, and so she is internally irrational. Assuming that she is neither ignorant nor compelled to believe as she does, she must deceive herself intentionally. Thus, Davidson maintains that to make sense of internal irrationality, we must postulate the possibility of there being mental causes that are not reasons for the beliefs they cause.[46] That is, there must be mental states capable of producing beliefs that do not serve as justifying reasons for those beliefs.

---

[44] Davidson (1985), 143. My emphasis.

[45] I will say more about the importance of the distinction between self-deception and wishful thinking in the next chapter. For our purposes in this chapter, it is sufficient to note at this point a) that "the assimilation of self-deception to wishful thinking fails to do justice to our vernacular conception of the phenomenon" (cf. Scott-Kakures, 1996, 140), and b) that Davidson himself considers this distinction important to his account of irrationality. It also makes room for weaker types of "irrational" belief-formation than straightforward self-deception, and if we take the latter to occupy a place on a broad spectrum of irrationality, the distinction between wishful thinking and self-deception appears quite appropriate.

[46] Note that in the discussion of Pears above, when we discussed the two types of motivations that might be relevant to self-deception (i.e., the preference on the part of *Cs* and the motivating factor behind *PCs*'s self-deceptive act), each of these may be said to correspond to a

More specifically, Davidson postulates the following conditions for an agent *A*'s being self-deceived regarding a proposition *p*: "*A* has evidence on the basis of which he believes that *p* is more apt to be true than its negation; the thought that *p*, or the thought that he ought rationally to believe *p*, motivates *A* to act in such a way as to cause himself to believe the negation of *p*."[47] The claim here is that what motivates an agent to self-deceptively form the belief that not-*p* is the co-existing belief about the likely truth of *p* itself (and, presumably, a desire that *p* not be the case, or some other motivational state that inclines the agent toward believing that not-*p*). Davidson claims that, to make sense of this, "[w]hat we must do is find a point in the sequence of mental states where there is a cause that is not a reason; a specific irrationality by the agent's own standards of rationality."[48] And it is for this reason that Davidson proposes partitioning the mind. Here, as elsewhere,[49] he argues that only if we conceptually divide the mind into sub-agential, quasi-autonomous structures can we explain how, within the space of one mind, one mental event can cause another without being a reason for it. Thus, he claims that there must be "boundaries between parts of the mind … somewhere between any (obviously) conflicting beliefs."[50] And the irrational step in self-deception, then, is "the drawing of the boundary that keeps the inconsistent beliefs apart."[51]

---

belief-desire pair that, if considered a cause of the self-deceptive belief, would represent a mental cause that is not a reason for the belief it causes.

[47] Ibid., 145.

[48] Ibid.

[49] See also Davidson (1982; 1998; 2004).

[50] Davidson (1985), 147.

[51] Ibid., 148. It appears to be the desire to avoid accepting the requirement of total evidence, on Davidson's account, that plays the causal role in the drawing of this boundary.

This results in what Pears calls a "functional" account of self-deception, as it is the causal role played by a particular belief that determines its being assigned to one system or another. In other words, a mental "schism" is postulated wherever intentional internal irrationality is to be explained: "For if someone is competent to avoid a piece of irrationality, the relevant cautionary belief will be somewhere within him, and, if it does not intervene and stop the irrationality, it will be assigned to a sub-system automatically by the function criterion."[52] This results in a view on which to say that two "parts" of the mind are distinct is to say that the mental states that constitute each part stand in tight logical relations with each other, but not with the mental states of the other part. Nevertheless, these parts can causally interact with one another, just as two people can interact. For example, my desire to make get you into my apartment may lead me to bake cookies, the smell of which cause you to form a desire for cookies that lead you to enter my apartment in search of baked goods. In such a case, my desire causally influenced your entering my apartment, but it does not serve as your reason for doing so. Similarly, Davidson argues, certain parts of one mind may causally interact with one another, without one part's motivation serving as a reason for another part's, say, forming a certain belief. This requires, of course, assuming that individual parts of the mind "must show a larger degree of consistency or rationality than is attributed to the whole."[53]

Thus, Davidson attempts to put forward an intentionalist account of self-deception similar to the stronger accounts we examined above. However, Davidson differs from Pears insofar as he insists that his mental partitions need not themselves represent separate centers of agency. His account likewise differs from Freud's insofar as he does not claim that the partitions of the mind must be separate and independent of one another.

---

[52] Cf. Pears (1984), 69.

[53] Davidson (2004), 181.

In contrast, he suggests that mental compartments themselves may "overlap" and share many of the same mental states and properties:

> I do not assume that the divisions are fixed, or that they deserve such names as conscience, courage, intellect, or id. More important, I do not think of the boundaries, however permanent or temporary, as separating autonomous territories. The territories overlap: there is a central core of mostly ordinary truths which the territories share. … Where territories differ is in the dissonant details. … Of course, this could not be the only difference: each of the contradictory beliefs needed a supporting phalanx of ideas. … The image I wished to invite was not, then, that of two minds each somehow able to act like an independent agent; the image is rather that of a single mind not wholly integrated; a brain suffering from a perhaps temporary self-inflicted lobotomy.[54]

For the above reason, Davidson shies away from giving a robust metaphysical account of these parts or divisions between parts. Rather, he claims that "such boundaries are not discovered by introspection; they are conceptual aids to the coherent description of genuine irrationalities."[55] Later, he writes: "I spoke of the mind as being *partitioned*, meaning no more than that a metaphorical wall separated the beliefs which, allowed into consciousness together, would destroy at least one."[56] Thus, for Davidson, although we need, conceptually, to divide the mind in order to be able to understand how internal irrationality may occur, we should not suppose that the parts proposed are themselves tiny homunculi, operating from their own centers of consciousness.

So how does Davidson's view fare in light of the above criteria that self-deception be intentional, internally irrational, and distinct from other "irrational" belief-forming processes like that wishful thinking? On the face of it, Davidson's account appears to account for all three criteria, insofar as he postulates a mental cause that is not a reason for what it causes by supposing the existence of explanatory mental partitions wherever

---

[54] Davidson (1998), 8. Cf. also Davidson (2004), 181, n.6.

[55] Davidson (1985) 147.

[56] Davidson (1998), 8.

an irrational belief transition occurs. As with the depth-psychological model, the static paradox does not appear to pose a significant threat to Davidson's view, insofar as a mental "schism" is postulated wherever the holding of two beliefs would, in theory, cause the agent to give up one or the other. The self-deceptive belief that not-*p* does not occupy the same inferentially integrated mental sub-space as the belief that the evidence warrants the conclusion that *p*. In the case of Agnes above, the belief that the evidence points to Ralph's cheating on her occupies a distinct space from the belief that he is faithful, thereby allowing Agnes to self-deceptively believe in Ralph's fidelity. Of course, it is the former belief, paired with a desire that he be faithful that motivates and eventually causes (or causally sustains) this latter, self-deceptive belief. But the awareness that this is so, too, is partitioned off and not directly accessible to the part of Agnes that believes he is, in fact, faithful.

Nevertheless, there are some serious worries that Davidson's account, too, must face. First, it is not entirely clear *how* these particular beliefs come to occupy separate parts of Agnes' mind. As we have seen, Davidson himself is hesitant to suppose that his conceptual partitions represent psychological realities, where one part plays the role of the perpetrator and the other of the victim of the deception. So we are not to take his postulations of mental partitions as literally as on the depth-psychological model. Yet it is still difficult to explain how the relevant beliefs become psychologically "separated" on this account. For if the agent knows what she is doing, the dynamic paradox threatens to undermine her ability to thwart the requirement of total evidence in favor of her preferred belief. But if she is unaware of the drawing of the schism that keeps her beliefs apart, she appears no better off than the ignorant agent who doesn't know better or the psychologically compelled agent who has no control over her beliefs. So the mystery remains: How does the self-deceiver who intends her deception "draw" this boundary between her inconsistent beliefs?

Toward the end of his discussion of weakness of will, Davidson himself admits that the weak-willed agent "cannot understand himself: he recognizes, in his own intentional behaviour, something essentially surd."[57] But this threatens to undercut Davidson's claim that the weak-willed agent's behavior is something he *does*—and does intentionally. Sarah Buss (1997) makes a similar point:

> It thus seems that, on Davidson's own account, the weak-willed agent cannot help but regard her action as something that just happens to her; she cannot help but feel like a passive bystander to her own behavior; her experience is indistinguishable from the experience of the *unfree* agent who finds herself compelled by an irresistible desire.[58]

And we might say the same thing about the self-deceiver. On Davidson's view, it appears that the self-deceiver ends up arriving at her self-deceptive belief "just like that," as it were. She cannot understand how she came to this belief, as the relevant self-deception-sustaining information regarding the belief's origin and/or justifiability is now inaccessible to her. She appears a "passive bystander" to her own deception. And thus it is unclear how the self-deceiver deceives herself *intentionally*.

There are also related worries about Davidson's requirement of internal irrationality in self-deception. We take an internally irrational agent to be someone who not only *ought* to know better, but someone who *does*, in fact, know better. However, as far as the self-deceived agent is concerned, she doesn't know better, so long as her relevant beliefs are partitioned off by her self-deceptive activity. Of course, the Davidsonian may retort that since the agent cannot adequately understand the origins of or justification for her self-deceptive belief, she will experience a degree of cognitive tension indicative of internal irrationality. And it is true that if she finds herself simply believing for no reason or "just like that," she may come to suspect that her belief has

---

[57] Davidson (1980), 42.

[58] Buss (1997), 30.

been irrationally formed. But if this is the case, then she is in a different epistemic situation than is presupposed by Davidson's account of self-deception. She appears either a stranger to herself, in the compelled sense described above, or she does have access to the epistemically relevant beliefs regarding her situation. Alternatively, she may invent (read: rationalize) a reason for her so believing, but if she is ignorant of the real cause of her belief, it does not seem so clear that she is internally irrational, in the sense Davidson wants.

This leads to a further worry about Davidson's requirement that an account of self-deception must distinguish between wishful thinking and self-deception proper. Davidson claims that the wishful thinker is not necessarily internally irrational, insofar as she believes something that, by her lights, is fully rational. But is not clear how the self-deceiver ends up being all that different from the wishful thinker. If the relevant beliefs are "exiled," such that the agent has no awareness of or access to them, then she, too, believes rationally by her own lights. Furthermore, even if we can understand how belief that is supposed to initiate the *intentional* causal process resulting in the acquisition of a belief to the contrary becomes inaccessible to the part of the mind deceived, it is unclear how such a belief could causally *sustain* the latter belief, if the former is no longer available to the agent. The false belief must be maintained *because* of the original belief to the contrary. However, it appears that if the original belief is inaccessible to the agent, the desire to hold the false belief must perform this task. And then it appears that following the initial (mysterious) self-deceptive step, the agent is not in a state of self-deception, but rather is engaging in mere wishful thinking. She no longer realizes that the evidence counts against the belief she currently maintains (since she no longer has access to the contrary belief). So the internal irrationality supposed to be present in self-deception is lacking. As Scott-Kakures notes:

> This would appear to recapitulate our earlier difficulties. The desire that not-p motivates me to turn away from my justified belief that p. But what the rogue desire accomplishes is this as

> well: It brings about, *de trop*, as a cause and not a reason, the exile of the requirement of total evidence, that epistemic canon which counsels me to believe on the basis of my best reasons. And if this is so, I do not realize that I come to believe in a way which violates my best standards of reasoning. For the violation of the requirement of total evidence has not been achieved with open eyes. We are left with wishful thinking. I come to believe for objectively bad reasons, but there is no internal irrationality.[59]

Pears, too, is concerned about Davidson's ability to distinguish between the various levels of irrationality. "If we adopt the functional theory," he writes, "we shall say that there is a schism whenever there is irrationality that the person is competent to avoid, even if it is a low degree of irrationality."[60] So long as the agent may be said to be competent to resist a particular piece of irrationality, Pears claims, the relevant cautionary belief must be present. But if it fails to intervene appropriately in the main system, Davidson's functional theory will automatically assign it to a sub-system, since it is the functioning of the belief itself that determines to which side of the schism the theory assigns it. "So the thesis, that no degree of avoidable irrationality is possible without a schism, will be true by the definition of the word 'schism'."[61] But now the distinction between wishful thinking and self-deception collapses, and Davidson is left with a phenomenon that looks even more mysterious than that with which he began.

Thus, in the end I am not sure that Davidson's account escapes many of the criticisms launched at the stronger divided-mind theories. Although his picture of self-deception is somewhat less metaphysically suspect than the neo-Freudian account, insofar as he requires that the self-deceived mind be divided into separate mental components, at least one of which must become opaque or inaccessible to the conscious

---

[59] Scott-Kakures (1996), 43.

[60] Pears, op. cit.

[61] Ibid., 70.

agent, there still appears to be a difficulty in locating the agent or giving her the mental control required to intentionally deceive herself.[62]

Likewise, the belief state transition from the belief that *p* to the belief that not-*p* (where the former belief is a cause but not a reason of the latter) is quite mysterious ~~(and does not appear to be under the control of the agent herself)~~. Although Davidson appears to maintain that this is the irrational step involved in engaging in self-deception, it seems that in order for such a transition to be possible, there must be *two* mental anomalies: a) a belief must cause a contrary belief without being a reason for it, and b) the agent must somehow "exile" the former belief from his explicit awareness, or somehow keeps these two beliefs separate. How either (a) or (b) actually occurs remains relatively mysterious. If (a) occurs prior to (b), it is not clear the agent can deceive herself intentionally. If, on the other hand, (b) occurs prior to (a), the question arises as to how the exiled belief can give rise to the contrary belief. Thus, it appears that even Davidson's weaker partitioned-mind theory cannot do the work it is supposed to do, for the agent no longer appears responsible for her deception, except insofar as she at one time mysteriously made an irrational "leap" from an undesirable belief to its contrary.

2.7 Conclusion

In this chapter, I have attempted to show that, although there are powerful reasons to put forward an intentionalist account of self-deception, the attempt to do so by partitioning the mind is not the best strategy for the intentionalist to adopt. Stronger depth-psychological models "solve" the epistemic difficulties at the cost of raising more worrisome metaphysical problems regarding the mind, the self, and agency in general. Weaker functional accounts give us little to no metaphysical picture whatsoever. And

---

[62] I mean here that if we cannot locate a central agent who exercises some degree of control over certain of her competing mental compartments, the agent (as a mere "amalgam" of these compartments) does not at all appear to be in control of her actions, including her self-deceptions.

both types of account fail to explain how self-deception is intentional and internally irrational.

The reader should understand here that I am not committed to denying the existence of mental compartmentalization. There is some good empirical reason to think that, at least occasionally, even "normal" agents experience a certain kind of mental partitioning.[63] What I am attempting to show is that such psychological divisions, even if real, do not provide the key to self-deception – at least if self-deception involves deceiving oneself intentionally and being somehow aware that the resulting belief strains one's own normative commitments. For this reason (and for others that will become clear in the next chapter), many philosophers claim that a better way to understand garden-variety self-deception is to adopt a view that does away with the strong analogy between interpersonal deception and self-deception. They propose that adopting a non-intentionalist view allows us to better make sense of the phenomena, while at the same time according with the results found in various empirical studies of irrational belief-forming processes. It is to a discussion of this type of theory that I now turn.

[63] Cf., for example, Gur & Sackeim (1979) and Quattrone and Tversky (1984).

CHAPTER 3. DEFLATIONARY ACCOUNTS: CAN WE MAKE SENSE OF SELF-DECEPTION WITHOUT INTENTION?

If divided-mind accounts of self-deception are unsuccessful in the ways we described in Chapter Two, what alternatives remain for the philosopher to explain the phenomenon? The failure of partitioned-mind theories to give a plausible account of intentional self-deception has led many philosophers to adopt the stance that the only way to avoid the static and dynamic paradoxes is to deny that there is a strong analogy between interpersonal deception and self-deception and claim that self-deception is largely *un*intentional.

Led by philosophers like Alfred R. Mele, this alternative conception of self-deception appears to be supported by our best theory of how the mind works—and about how human beings actually reason. According to this conception, most, if not all, instances of self-deception are reducible to some kind of motivational biasing, which occurs independently of anything the agent does intentionally.[1] In general, non-intentionalists claim that their theory is superior to intentionalist theories in two related ways. First, they claim that non-intentionalist theories can explain the same phenomena with fewer controversial commitments. Non-intentionalists charge that intentionalist accounts contain unnecessary metaphysical baggage, or that they mistakenly attribute too much deliberate and explicit reasoning to the self-deceiver. Second, non-intentionalists claim that their theories more plausibly explain the empirical data.[2] [3] The burden of

---

[1] See Johnston (1988) and Barnes (1997) for other examples of non-intentionalist theories of self-deception.

[2] By 'empirical data' here, I refer both to our everyday observations of what we take to be self-deceptive behavior and to the plethora of empirical studies in psychology, neuroscience, cognitive science, anthropology, and other fields that purport to examine certain kinds of belief formation (e.g., irrational, motivated, or otherwise biased belief).

[3] Whether these two claims really are separate is not clear. Quite often the claim that non-intentionalist theories are explanatorily superior to intentionalist theories depends solely on the fact that the former are "simpler" (cf., e.g., Mele, 1997, 96). Others, however, appear to claim

proof is thus placed on the "traditionalist"[4] about self-deception to show that self-deception plausibly mirrors interpersonal deception while at the same time accounting for the empirical data and our commonsense intuitions regarding the phenomenon.

In this chapter, I will examine the general approach non-intentionalists take to explaining self-deception and raise some worries about taking such a limited approach. I intend to show that deflationary explanations generally fail to account for certain crucial features of self-deception and self-deceptive behavior, and that a traditionalist account is in much better shape to make sense of these features. I hope thereby to at least lighten the burden on the traditionalist and pave the way for an alternative approach to self-deception.

3.1 Non-intentionalist accounts of self-deception: Mele's four conditions.

Alfred Mele has proposed that "the claim is *unwarranted*, [though] *not* incoherent, that intentions to deceive ourselves, or intentions to produce or sustain certain beliefs in ourselves…are at work in ordinary self-deception."[5] Thus, although he allows for the possibility of a certain kind of intentional self-deception (e.g., intentionally falsifying one's diary on the assumption that one will later forget one has done so and thus be "taken in" by one's own lies), he claims that we can better make sense of

---

that we cannot make sense of the empirical data at all, if we claim that self-deception is necessarily intentional (cf., e.g., Barnes, 1997, 95).

[4] The reader will notice that I often shift to talk of "traditionalist" vs. "deflationary" accounts of self-deception in this chapter. For my purposes, I use the former term to designate those accounts that attempt to maintain a close analogy between self- and other-deception. Intentionalist theories generally fall under this rubric, but there are also philosophers of traditionalist persuasions, who do not view themselves as straightforward intentionalists (e.g., Dion Scott-Kakures). Deflationary accounts, on the other hand, are almost exclusively comprised of non-intentionalist theories.

[5] Mele (1997) 99.

"garden-variety" cases of self-deception[6] by appealing to certain motivational biases, which play a significant causal role in the generation or sustenance of a particular false belief in the self-deceived agent.

Mele concedes that the appeal to motivation does not explain all cases of irrational belief formation. Indeed, most cases of irrational believing arise due to non-motivational factors. In an oft-cited work, Nisbett and Ross (1980) discuss in great detail the pervasiveness of inferential mistakes and the biased treatment of evidence in human agents:

> [I]nferential and judgmental errors arise primarily from nonmotivational—perceptual and cognitive—sources. Such errors…are almost inevitable products of human information-processing strategies. In ordinary social experience, people often look for the wrong data, often see the wrong data, often retain the wrong data, often weight the data improperly, often fail to ask the correct questions of the data, and often make the wrong inferences on the basis of their understanding of the data. With so many errors on the cognitive side, it is often redundant and unparsimonious to look also for motivational errors. We argue that many phenomena generally regarded as motivational…can be understood better as products of relatively passionless information-processing errors than of deep-seated motivational forces.[7]

Mele notes that these so-called "cold" (i.e., unmotivated) biases (e.g., confirmation bias, and biases tied to the vividness of information, the availability heuristic, the tendency to search for causal explanations, and so on) do very often affect our cognition in ways that may result in an agent's holding an unwarranted belief.[8] However, he claims that, in general, the state in which the self-deceiver finds herself is

---

[6] Many claims regarding what counts as a "garden-variety" case of self-deception (and what does not) in the literature are either question-begging or an attempt at intuition-pumping. But in the end, whether a particular phenomenon is of the "garden variety" may be purely an empirical matter. However, to avoid begging the question against the non-intentionalist, I will attempt to employ examples that non-intentionalists themselves would accept as being of the "garden variety."

[7] Nisbett & Ross (1980), 12.

[8] Cf. Mele, op. cit., 93-4; Mele (1987), 144-5.

not solely a product of the activity of these cold biasing mechanisms. Rather, he maintains that mere cold biasing is distinct from self-deception. In particular, motivational states (e.g., wishing or desiring that *p*) often lead us to treat evidence in a way that "can prime mechanisms for the cold biasing of data in us without our being aware, or believing, that our evidence favors a certain proposition."[9] For example, we may negatively or positively misinterpret data regarding *p*; we may selectively attend to certain features of a situation that support *p*, while failing to attend to features that support not-*p*; we may even more actively gather evidence for *p*, while overlooking evidence against *p*. And when these "hot" (motivationally biased) strategies are added to one or more of the aforementioned "cold" (cognitively-biased) mechanisms, Mele thinks it is not difficult to explain how it is that agents are regularly able to deceive themselves. It is therefore not surprising that, for Mele, et. al, there is no strong analogy between interpersonal deception and self-deception, except insofar as we often speak of a person who is in error with regard to *p* as being "deceived" about *p*.[10]

Mele puts forward the following four conditions as *jointly sufficient* for an agent *S*'s entering self-deception in acquiring a belief that *p*:[11]

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*.

---

[9] Mele (1997) 95. Note that this claim is significantly weaker than the claim made by Davidson, et. al., that the agent who is self-deceived in believing that *p* must have some awareness that the evidence favors not-*p*. I discuss this point in more detail below.

[10] Ibid. 92. Mele also suggests that there is also a natural use of 'deceive' in the active voice, namely 'to cause to believe what is false'—and that this need not be intentional. Although I take issue with such a use of the word, if we accept this lexical usage, then there still may be some room for some analogy between interpersonal deception and self-deception, just not the one indicated by the most common usage of the term.

[11] Mele claims that similar conditions can be given for cases in which an agent enters self-deception in *retaining* a belief (cf. Mele, 1987, 131-2). I challenge this claim in detail below.

4. The body of data possessed by *S* at the same time provides greater warrant for ~*p* than for *p*.[12]

Condition 1 makes the lexical claim that the belief the agent acquires be false. Otherwise, Mele argues, the agent would not count as *deceived*.[13] Of course, he does not deny that one may come to believe a proposition in motivationally biased ways that, despite the prevalence of evidence against it, turns out to be true. However, he argues that whereas one may be self-deceived in acquiring the true belief that *p on the basis of e* (where *e* represents evidence that, in this case, does not warrant the believe that *p*), one is *not* self-deceived "in acquiring the belief that *p simpliciter*"—and it is the belief that *p simpliciter* that interests Mele.[14] Thus, he denies that one can enter the *condition* of self-deception in believing a proposition unless it is, in fact, false.[15]

Condition 2 paves the way for Mele's non-intentionalist account. I take the occurrence of some sort of "hot" biasing to be central to Mele's account of self-deception. Indeed, it seems implausible to think that an agent lacking either component of Condition 2, namely a relevant motivation to believe that *p* or a biased treatment of the evidence for *p*, would count as self-deceived.[16]

---

[12] Mele (2001), 50-1. See also Mele (1997), 95, and Mele (1987), 127.

[13] Mele (2001) 50-1.

[14] Mele (1987) 128. See pp.127-8 for a further demarcation of being deceived *in* believing a certain proposition and being deceived *into* believing that proposition.

[15] This is a claim I will go on to deny in Chapter Four, but this denial rests on my different understanding of what constitutes the "condition" of self-deception. I think Mele and I agree that the falsity of a proposition does not, by itself, affect the *dynamics* of self-deception (cf. Mele, 1997, 95). For the claim that both interpersonal deception and self-deception can involve the victim's coming to believe a true proposition, see Barnes (1997) 8-9, 54.

[16] In Chapter Five of Mele (2001), he leaves room for the possibility that emotional biasing might play the role here occupied by motivational biasing, but it is not clear to me that these two categories are really all that distinct. For an account that reduces self-deception to a kind of emotional biasing, see Baljinder (2003).

Condition 3 is employed to rule out "inappropriate" causal chains, in which an agent's motivation enters into the causal chain of his believing a certain proposition in a deviant way. To give an example similar to Mele's, suppose my desire to be esteemed by my colleagues causes me to become so nervous during a public lecture that I trip over a piece of loose carpet and bump my head on the chalkboard; suppose also that the injury just happens to produce the false belief in me that my colleagues respect me. Here, we have a case that appears to meet conditions 1, 2, and 4, but we would not likely be inclined to call this a case of self-deception because my false belief, while caused by my desire, is brought about via a deviant causal chain. Of course, determining which causal roles count as "proper" or "deviant" is difficult, but Mele seems right to assume that an adequate account of self-deception relies on such a distinction.[17]

Finally, Condition 4 claims that the self-deceived agent believes against the weight of the evidence she possesses. Mele argues that this condition is not a *necessary* condition of self-deception. There may be cases in which, due to, e.g., selective-evidence gathering, the body of evidence *S takes herself to have* at the time of entering into self-deception favors the truth of *p* over that of ~*p*, even though there is more evidence for ~*p* readily available to *S*, which she overlooks, due to the operation of the motivational bias in question. Nonetheless, Mele claims, in most cases of ordinary self-deception, self-deceivers do, in fact, believe against the evidence "readily available to them."[18] It is also important to note that, on this account, Condition 4 does not require that *S* be *aware* of the fact that she believes against the evidence.[19]

---

[17] Much of the motivation for Condition 3 appears to stem from Mele's claim that the self-deceived agent is somehow competent to avoid her deception. I discuss this claim in more detail below.

[18] Mele (1997), 95.

[19] Ibid., 102, n.16. This point will become important later on.

Therefore, Mele claims that normal, "garden-variety" instances of self-deception do not need to be modeled on intentional action, in order for us to be able to make sense of such behavior. Indeed, if we accept the above conditions as jointly sufficient for self-deception, we do not appear to encounter the aforementioned paradoxes that make self-deception appear so problematic. At no time in the acquisition of her self-deceptive belief, need the agent hold contradictory beliefs, so the static paradox is thereby avoided. Neither do we appear to encounter the strategic paradox, since on this account there is no intentional strategy the agent employs in order to deceive herself.

3.2 Initial Objections to Mele and Some Possible Responses

These apparent advantages notwithstanding, one might be inclined to challenge the view on a few counts. First, one might worry that Mele's deflationary account fails to appropriately distinguish between self-deceived agents and either epistemically ignorant agents or psychologically compelled agents.[20] Second, one might object that Mele fails to make an important conceptual distinction between what philosophers and psychologists call "wishful thinking" and self-deception proper. Third, self-deceived agents appear to engage in all sorts of intentional behavior (e.g., intentionally focusing on positive thoughts and avoiding negative ones), so how is it that self-deception is *unintentional* on Mele's account? I will examine each of these initial objections in turn.

3.21 On the distinction between self-deception and cases of ignorance or compulsion.

According to the first objection, Mele's account cannot appropriately distinguish between self-deceived agents and either merely ignorant or mentally-compelled agents. Let us first turn to the problem of ignorance. If the self-deceived agent is not aware that the evidence points to her belief's being false (or that the cause of her belief is some

[20] Mele is not a "skeptic" about self-deception in the senses I employed in Chapter One. Thus, he must be able to distinguish his account from these two types of skeptical accounts.

motivational bias, as opposed to a rational sensitivity to the evidence for her belief), it appears that she is merely *ignorant* of some relevant fact about herself or her epistemic situation that precludes her holding the relevant true belief. In response to this objection, Mele claims that there *is* a relevant difference between the merely ignorant agent and the self-deceived agent, namely that the latter's belief is predominantly caused by a desire (e.g., that things be as she believes them to be) or by some other strong motivational attitude (e.g., fear, jealousy, self-loathing, etc.).[21] The former's belief may be the result of straightforwardly poor reasoning on the part of the agent—or even by good reasoning, in cases where the agent unknowingly lacks the relevant evidence that would cause him to acquire a more justified belief. But, according to Mele, the formation of the ignorant agent's false belief is not *motivated* in the way that the self-deceived agent's is. In other words, self-deception is more than just being "mistaken."

However, we might then wonder how such an account can distinguish between self-deceived agents and those who are psychologically compelled to believe as they do. In cases of psychological compulsion (e.g., in obsessive-compulsive disorder), an agent may simply find herself entertaining a recurring thought or belief, the pull of which she is unable to resist. She cannot "help herself" in so believing (and thus in acting accordingly). But if self-deception is reducible to the mental causation of a belief by some predominant motivation, how does the self-deceived agent differ from the obsessive-compulsive agent? Here, Mele claims that not only is self-deception more than simply being in error regarding a proposition, it is also something the agent is competent to avoid.[22] The self-deceived agent, as opposed to the obsessive-compulsive agent, has it "within her power" to take psychological measures to avoid being motivationally biased.

---

[21] Mele (1987), 123.

[22] See, for example, Mele's (1987) discussion on self-control in the case of akratic belief (116-7).

Thus, Mele claims, the self-deceived agent is neither merely "accidentally" deceived, nor is she, strictly speaking, compelled to believe as she does.[23]

3.22 On the relationship between wishful thinking and self-deception.

Regarding the second objection, i.e., that his account of self-deception fails to distinguish between wishful thinking and self-deception proper, Mele is perfectly content to admit that wishful thinking may be a species of self-deception.[24] On this view, wishful thinking may be demarcated from other cases of self-deception by noting that in wishful thinking, the believer is motivated by a *desire* (or "wish") to believe, whereas all sorts of other motivational (or emotional) states may be responsible for the formation of self-deceptive beliefs. However, he also claims that "the difference may lie in the relative strength of relevant evidence against the believed proposition: wishful thinkers may encounter weaker counterevidence than self-deceivers."[25] That is, wishful thinkers may be in a better position with regard to the evidence than self-deceivers, despite the fact that the mechanisms at work in the production of the false belief are more or less the same. For example, self-deceivers may be more resistant to negative evidence than their wishfully thinking counterparts.[26]

---

[23] Barnes (1997) makes a similar claim: "[A] self-deceptive belief is always other than a compulsive belief. … Self-deceptive belief is, I contend, neither compulsive belief nor intentional belief" (46, n.29). Furthermore, although she claims that one might not be able to resist being biased in certain cases, one can "be on the lookout for bias," thus making it more difficult for one to become self-deceived. Furthermore, she argues that a charge of epistemic irresponsibility or neglect may be legitimately levied if one does not "do all that one should to ensure that one's beliefs are true…to take steps to correct one's shortcomings in belief acquisition by disciplining one's mental habits…" (84). This implies that she takes self-deception to be, in general, resistible.

[24] Mele (1987) 135.

[25] Mele (1997) 100.

[26] Of course, where Mele wants to draw the line between wishful thinking and self-deception so viewed is unclear, especially given that they arise from the same sorts of psychological processes.

3.23 On the intentional behavior of self-deceivers.

Finally, in regard to the third objection, Mele claims that agents may undertake all sorts of intentional behavior with the result that they become self-deceived, but that this does not entail that the deception *itself* is intentional: "Intentional cognitive activities that contribute even in a relatively straightforward way to self-deception need not be guided by an intention to deceive oneself."[27] For example, one may intentionally focus on pleasant thoughts (or intentionally avoid unpleasant thoughts), and this selective focus may causally contribute to one's current or future self-deception, but this does not entail that one *intends to deceive oneself.* According to Mele, it is rarely (if ever) the case that one intends one's deceit. Indeed, the content of one's motivation for self-deceptively believing that *p* is not generally, say, the desire that one believe that *p*, but rather that *p* actually be true. And such a desire can causally contribute to one's believing that *p* is true without a concurrent intention to so believe.

Whether at the end of the day Mele's responses to these initial objections are satisfactory, we shall examine later in this chapter. I am inclined to think that they are not. For now, however, it is sufficient to note that Mele's responses are, at the very least, initially plausible and appear to place the burden on the traditionalist to show that her account has distinct advantages over the deflationary account, especially given that it seems the non-intentionalist has the upper hand in terms of explanatory parsimony. So how might the traditionalist attempt to counter such a position?

3.3 Traditionalist arguments against the non-intentionalist position.

A promising strategy for the traditionalist regarding self-deception is to show that the "simpler," "more parsimonious" account offered by non-intentionalists cannot account for certain crucial features of self-deception. Indeed, if it turns out that there are

---

[27] Ibid., 98.

phenomena that cannot be adequately explained on non-intentionalist pictures of self-deception, then this will leave room for the traditionalist (including the intentionalist traditionalist) to get his foot in the door. It is to a discussion of potential candidates for this kind of strategy that I now turn.

3.31 The Presence of Dual-Beliefs in Self-Deceivers.

One popular traditionalist attempt to challenge the non-intentionalist is to meet the so-called "dual-belief requirement" (hereafter, *DBR*), i.e., to show that self-deceived agents actually do simultaneously believe contradictory propositions. Mele claims that the burden of proof is on the intentionalist (or any traditionalist about self-deception, for that matter) to show that self-deceived agents typically hold contradictory beliefs.[28] If this can be shown to be the case, many traditionalists think, then deflationists regarding self-deception must drastically revise their accounts to include this phenomenon.

Here it is important to note that Mele does not deny that agents may sometimes have contradictory beliefs. And this seems right. It is often the case that an agent holds contradictory beliefs, which are separated by enough "cognitive distance" that she consistently fails to notice that she believes a contradiction.[29] Likewise, it is very common for an agent to believe that *p* and that *q* and yet to be completely unaware that the former belief commits her to believing that *r*, whereas the latter commits her to believing that not-*r*.[30]

---

[28] See, for example, Ibid., 101, 128 and Mele (2001), 76-93.

[29] In other words, an agent may dispositionally believe *p* in certain circumstances (or under some description) and not-*p* in different circumstances (or under a different description), without thereby being aware that she believes a contradiction.

[30] Cp. my example in Chapter One of first-year philosophy students who maintain a) that moral relativism (oftentimes in the form of cultural relativism) is true and b) that we ought, objectively speaking, to be tolerant of other cultures' moral beliefs and practices. In such cases, the student is often unaware that holding (a) implies that there are no objective moral truths and that holding (b) implies that there is at least one objective moral truth.

However in both of these kinds of cases, there is some form of ignorance that may explain what would otherwise represent an irrational doxastic commitment to contradictory propositions. In cases of the former type, the agent is unaware that two or more of her beliefs conflict. Perhaps they always arise in completely different contexts that rarely if ever overlap or occur in close temporal relation to one another.[31] Here, we can easily understand how some agents would fail to notice this conflict. In cases of the latter type, the agent fails to make certain relevant inferences, which themselves would come into conflict.[32] And while a failure to make certain relevant inferences may turn out to be a characteristic of some instances of self-deception, it is unlikely to represent a genuine case of self-deception if the agent's motivations or reasons play no significant role in explaining why she fails to make these particular inferences. It may simply be a result of poor (but largely unmotivated) reasoning. But Mele's challenge to the traditionalist is to find plausible cases of *self-deception* in which agents simultaneously believe a contradiction. Can such a challenge be met? Do we have reason to think that self-deceived agents can and do typically hold contradictory beliefs?

---

[31] A graduate student may, for example, believe herself to be outgoing (where the content of her belief is simply "I am an outgoing person"), but she may find herself entertaining this belief only when amongst friends. At professional conferences, however, she may feel rather introverted and may thus entertain the simple belief that she is not an outgoing person. But these situations may be separated by enough cognitive and temporal distance that the graduate student never realizes that she believes both that she is an outgoing person and that she is not. (Of course, if she does realize this, she may revise her beliefs to represent this fact: e.g., "I am outgoing amongst friends but not amongst strangers.")

[32] The first-year philosophy student may make this mistake when he professes honestly that he believes that there are no absolute moral truths and yet sincerely asserts that we all ought (in a moral cognitivist, absolutist sense) to be tolerant of other cultures' moral practices. He simply does not see that committing himself to the latter precludes him from consistently asserting the former. Yet we may (at least somewhat plausibly) claim that he really does have these two propositional beliefs. He merely fails to draw the inferences that these beliefs commit him to.

Neil Levy (2008) cites cases of "anosognosic hemiplegia," the persistent denial of partial-paralysis by sufferers.[33] He claims there is good empirical evidence that such patients do hold contradictory beliefs in the sense required by *DBR*. For example, cold water poured into the left ear of such patients can temporarily relieve their anosognosia (presumably by stimulating the parts in the right hemisphere of the brain that are active in detecting anomalies), with the result that the patient will report recognizing he has been paralyzed *ever since his stroke*. This may indicate that the belief that she is paralyzed on one side of her body has been present in her all along. Likewise, anosognosic patients deny their condition more vehemently than do non-anosognosics, avoid undertaking tasks requiring use of the paralyzed body part, and attempt to rationalize away failures to perform tasks with the paralyzed part on command—e.g., "I'm too tired," or "I don't feel like it," or even "That's not my arm, it's my mother's. She's hiding under the table."[34] Furthermore, the denied "knowledge" that she is partially paralyzed appears to be "dispositionally available" to the anosognosic, if prodded enough.[35] However, the repeated attempts by anosognosics to undertake tasks requiring the use of their paralyzed limbs suggest that these patients are, indeed, sincere in their avowal that they are not paralyzed.

If this is correct, it appears possible that agents can and do simultaneously occupy conflicting doxastic states that are (at least sometimes) accessible to consciousness without their rejecting one or the other of the beliefs. Levy goes on to argue (rather persuasively) that mere neurological causes for anosognosia are insufficient to explain the phenomenon, and that there must be a motivational element present as well to explain

---

[33] Levy (2008).

[34] I take this last example from V.S. Ramachandran's fascinating interview with Errol Morris on anosognosia on the New York Times' Opinionator Blog. Cf. Morris (2010).

[35] Levy (2008), 10.

why some hemiplegics become anosognosic and others not. Thus, he concludes that most instances of anosognosic hemiplegia are genuine cases of self-deception, and that this "places the burden of proof squarely back upon the shoulders of the deflationists. No longer can they argue that their view is less psychologically extravagant than that of their opponents."[36]

However, even if we conclude with Levy that "doxastic conflict…is a real feature of human psychology,"[37] this still does not show that self-deceived agents *simultaneously believe* p and not-p, and it certainly does not demonstrate that they intend their deception. If anything, it merely shows that self-deceived agents may experience conflicting doxastic states, none of which must amount to a full-blown belief. Levy himself appears to backtrack from *DBR* a bit by often referring to "strong suspicion" instead of "belief." But Mele's challenge to the traditionalist was to show that self-deceived agents hold contradictory *beliefs***.**

It appears, then, that the traditionalist regarding self-deception is going to have trouble putting forward a view that meets *DBR* without invoking the kind of problematic partitioned-mind theories discussed in the last chapter. This suggests that *DBR* is too strong a requirement. Although it may be the case that some level of cognitive dissonance or psychic tension is characteristic of self-deception (as I will discuss later), this in no way implies that the self-deceived agent must straightforwardly believe contradictory propositions. It thus remains open to a traditionalist of the intentionalist persuasion to put forward an account of self-deception that remains relevantly analogous to interpersonal deception while at the same time rejecting the strong form of *DBR*, which requires the agent to fully believe both *p* and not-*p*.[38]

---

[36] Ibid., 13.

[37] Ibid.

[38] I discuss this possibility further in the next chapter.

The real challenge to the traditionalist is to show that the self-deceived agent *knows* or *is aware* on some level that she occupies conflicting states in a sense that would require an attribution to the agent of a kind of intentional collusion in her own cognitive dissonance. Thus, the traditionalist might stipulate that the agent *suspect* or have some sort of *minimal awareness* of the contradictory or inconsistent nature of her doxastic commitments—or, at the bare minimum, that this awareness must be at least *potentially available to consciousness*.[39]

3.32 Internal Irrationality, Psychic Tension, and Cognitive Dissonance

This points to a more promising approach for the traditionalist to take against the deflationist. We can utilize one of the strategies we employed to motivate (and eventually reject) divided-mind accounts in the last chapter—namely, the appeal to internal irrationality.[40] As we noted in Chapter Two, the notion of internal irrationality seems to be central to any rigorous account of self-deception. However, we should examine this claim in a little more detail here, for unlike partitioned-mind theorists, non-intentionalists do not, in general, appear to take internal irrationality to be a genuine requirement for self-deception.[41] They identify the irrationality of self-deception with the fact that some motivational factor (as opposed to, e.g., a rational sensitivity to the evidence) plays a central role in bringing about or causally sustaining a particular false belief, or with the

---

[39] This rules out strong partitioned-mind views of the type discussed in the previous chapter. As we saw in Chapter Two, dividing the mind into quasi-agential subsystems appears to raise more problems then it solves. Of course, one might just claim that one of the doxastic states is conscious while the other is unconscious, without making any reference to "person-like" mental substructures. However, I am inclined to think that any strategy which attempts to make one or more of the relevant beliefs entirely inaccessible to consciousness can neither be said to meet *DBR*, nor our revised challenge. Cf. Hirstein (2000, 2005) for a similar position.

[40] Or, as Barnes (1997) calls it, "deep epistemic irrationality" (25).

[41] For example, Barnes (1997) claims that "although self-deceivers are always epistemically irrational, they are not…*deeply* epistemically irrational" (137, my emphasis).

fact that this belief falls short of some third-personal standard of rationality. Mele, for example, often refers to what he calls the "impartial-observer test" to determine whether an agent meets a necessary condition for self-deception. He writes:

> [If] *S* is self-deceived in believing *p*, and *D* is the collection of relevant data readily available to *S*, then if *D* were made readily available to *S*'s impartial cognitive peers (including merely hypothetical people), those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true.[42]

Mele argues that if *S*'s "impartial cognitive peers" (minimally, those who have neither the desire that *p* nor that ~*p* and who do not prefer avoiding being in error regarding *p* over being in error regarding ~*p*, and vice versa) would generally come to the same belief as *S* (i.e., the belief that *p*), then this may serve as an indicator that the agent, *S*, is not self-deceived. But the test also appears to point out a criterion one might adopt to explain what it is that makes one's self-deception regarding *p* irrational—not only is it biased, but it is biased in such a way that the vast majority of one's impartial cognitive peers would conclude that *p* is false.

I do not wish to go into all the worries I have about Mele's impartial-observer test here, but it is important to note that the necessary condition for self-deception constituted by failure of the test makes no reference to any sort of *internal* irrationality of the part of the self-deceived agent. She need not believe or be otherwise aware that her impartial cognitive peers would not believe as she does to count as self-deceived. In fact, Mele's account appears to rely heavily on the self-deceived agent's being *un*aware of this fact.[43] So what makes self-deception internally irrational on Mele's account? A plausible candidate for a "requirement" of internal irrationality might be Condition 4, according to

---

[42] Mele (2001), 106.

[43] If the agent *were* aware of this fact, it would likely cause her self-deception to become unstable in a way that might well lead to its demise, insofar as it might lead her to rationally revise her belief.

which the agent believes against the weight of the evidence she currently possesses—but Mele does not take this condition to represent a necessary condition for self-deception, and neither does he require that the agent *know* she believes against the weight of the evidence. Thus, by the self-deceiver's lights, she believes on the weight of the evidence. She is not internally irrational. Another plausible candidate for internal irrationality might be Condition 2, which states that the self-deceiver treats data (seemingly) relevant to the truth of the self-deceptive belief in a motivationally biased manner. But here, too, there is no requirement that the agent realize that she is treating the relevant data in motivationally biased ways. Again, by her lights, she is perfectly objective. Thus, there appears to be no room for internal irrationality on Mele's picture.[44]

But why claim that internal irrationality is necessary for self-deception in the first place? Deflationists like Mele claim that such a requirement is what has traditionally marred the efforts of traditionalists to give a non-problematic account of self-deception**.** And in the former's attempt to put forward an explanatorily parsimonious view, it is no surprise that they would jettison any requirement of internal irrationality from their accounts. However, there are a few reasons to think that to forego such a requirement would be to fail to acknowledge a very real kind of phenomenon.

One way to defend the claim that internal irrationality is a distinctive feature of self-deception is to again appeal to the proposed distinction between wishful thinking and self-deception. We saw that deflationists like Mele have no problem reducing self-deception to a kind of wishful thinking or vice versa. And it is true that we do speak loosely of people as "fooling" or "kidding" themselves when we think they are simply believing what they, e.g., want or are biased to believe. However, when we look at the

---

[44] In personal correspondence with Mele, he admits that he takes all instances of self-deception to be epistemically irrational, but he does not require that they be *internally* irrational—i.e., he does not hold that the agent be cognizantly irrational *by her own lights*.

kinds of attributions we make when we accuse people (or even our own past selves) of having been truly *self-deceived*, as opposed to merely biased, we notice that we tend to accuse them of somehow actually having known (or having been otherwise aware) of what it was they were doing. We distinguish agents who we think simply "ought to know better" from agents we think *do* know better—and we recognize that the latter kind of irrationality differs from, and is more robust than, the former kind. As Dion Scott-Kakures (1996) points out, "We…recognize that the appeal to wishful thinking is a far less drastic explanatory maneuver than is the appeal to self-deception."[45] Thus it appears that we do make a distinction between two kinds of phenomena in our everyday attributions of irrationality to certain agents—and it seems only natural to think that this distinction (i.e., between "shallow" and "deep" epistemic irrationality) corresponds to the distinction between wishful thinking and self-deception.

However, this is primarily a lexical point. And we do not want to rest the whole of our argument on the mere folk psychological usage of the term 'self-deception'. We want to show that there may be good reason to suppose that there *is*, in fact, such internal irrationality at work in cases we are inclined to call self-deceptive. One means of doing so is to appeal to the phenomenology of self-deception. What characteristics do self-deceived agents tend to exhibit? What do we notice about ourselves when we reflect on our own (past) self-deceptions?[46] As we have already seen in Chapter Two, many theorists of the intentionalist persuasion point to the cognitive dissonance typically exhibited by individuals we take to be self-deceived. In general, behavior we take to be

---

[45] Scott-Kakures (1996), 37.

[46] Some philosophers will worry that reflections on our own past self-deceptions may be skewed, re- or misinterpreted, or even themselves the products of self-deception or wishful thinking. This is, of course, a genuine worry. However, that does not mean we must discard all of our attempts at a phenomenological investigation of self-deception. We must merely approach them with a degree of concern and watchfulness.

indicative of self-deception points to an agent being somehow "in conflict" or "at odds" with herself in ways that rational, epistemically ignorant, or non-rational actors are not. There are two ways in which cognitive dissonance might be supposed to feature in self-deception. First, cognitive dissonance may serve as a *motivational* factor in producing self-deception. Second, dissonance appears to be a characteristic of the process of self-deception itself.

3.321 Cognitive Dissonance and Self-Deception.

The first possibility claims that an agent's motivation for her self-deception may rest in a kind of psychic tension, and we do in fact see such instances of self-deceptive motivation in real cases. Ziva Kunda (1990) notes that when agents find themselves with seemingly contradictory cognitions, or with cognized information that appears to conflict with their self-images, they are sometimes motivated to reduce the dissonance, e.g., by altering their attitudes (as in so-called "sour grapes" scenarios) or by acquiring certain false beliefs (as with the motivated overestimation of one's skills).[47] Of particular interest here is the claim put forward by several dissonance theorists that "dissonance arousal requires a threat to the self."[48] Although Kunda himself considers that there may be sources of cognitive dissonance other than perceived threats to the self, what is noteworthy is the possibility that self-deception is characteristically a *response* to a kind of cognitive tension or instability within the self.

Barnes (1997), following Johnston (1988), puts forward a similar view. Both proponents of non-intentionalist theories of self-deception, Barnes and Johnston claim that what plays the pivotal causal and motivational role in self-deception is a so-called

---

[47] Kunda (1990).

[48] Ibid., 484. I return to this notion in Chapter Four.

anxious desire.[49] For Barnes, a person has an anxious desire that *q*, "just in case the person both desires that *q* [for its own sake] and is anxious that it is not the case that *q*."[50] Self-deceptive belief, then, functions (in theory) to reduce anxiety, and self-deception is an agent's non-intentional response to one or more of her anxious desires. Such a view appears compatible with cognitive dissonance theory as we described it above. Here, it is the cognitive tension represented in the form of an anxious desire that gives rise to self-deception. Thus, whereas Mele has problems incorporating cognitive dissonance into his theory in the first place, Barnes is able to find a role for this kind of tension in generating self-deception.

For example, suppose that Jerry sees himself as a good husband (and views being a good husband as a desirable quality in a married man), yet he finds himself desiring a woman, Elaine, who is not his wife. Even if neither Jerry's belief about himself nor his desire for Elaine is deeply reflected upon, it would be no surprise to find that the presence of this belief about himself when paired with such a desire causes him to experience a certain cognitive unease. In Barnes's terms, the "anxious desire" in question might be Jerry's implicit desire to be a good husband, paired with his fear that his desire for Elaine suggests that he is *not* a good husband.[51] We can also easily imagine that this anxious desire is precisely what causes Jerry to form the conscious belief that he does not desire Elaine. Barnes would argue that this may all happen on the pre-reflective level, without

---

[49] Barnes departs from Johnston in claiming that self-deception is not always a species of *wishful* belief, though she agrees that self-deception is always a response to an anxious desire. She also takes issue with Johnston's claim that the self-deceiver need always recognize that the totality of her evidence points to not-*p* when she comes, self-deceptively, to believe that *p*. Cf. Barnes, op. cit., 32-3.

[50] Ibid. 38-9.

[51] Alternatively (though perhaps less plausibly), we might take Jerry's anxious desire to be the desire to have true beliefs about himself, paired with the worry that he has a false belief about himself, namely, that he is a good husband.

Jerry's having any sort of intention to deceive himself. Thus, we might take Jerry's case to represent a paradigm instance of non-intentional self-deception that is *motivated* by a kind of cognitive dissonance—in this case, by Jerry's anxious desire. And here, Barnes would argue, the self-deception itself serves the function of alleviating that anxiety.

However, despite the initial plausibility of the above case, a few worries loom. First, it is unlikely that Jerry would have experienced significant cognitive tension at all, if he had not, in some sense, believed (or suspected, or otherwise considered) that he desired Elaine and that he takes this to count as evidence against his being a good husband. Otherwise he would not experience the fear associated with his anxious desire in the first place. Had he been completely unaware of the relevant data—and, more importantly, of what he takes that data to entail—it would seem rather strange that he would be conflicted at all, for there would then be nothing motivating his fear that he is not a good husband. Indeed, it seems that the generation of the anxious desire requires Jerry's seeing the data *as evidence against* the proposition that he is a good husband and this seems to introduce a level of awareness that ought to make the non-intentionalist uncomfortable. For now, it is sufficient to note that if we suppose that cognitive dissonance in part explains *why* agents sometimes deceive themselves, we may already be implicitly importing a level of reflection on the part of the agent that the non-intentionalist would likely wish to deny.[52]

Nevertheless, if we accept Barnes' account, we see one way in which non-intentionalists may incorporate cognitive dissonance into their theory of self-deception, namely as a motivating factor that may set off the biasing mechanism in question, resulting in the agent's becoming self-deceived. However, it cannot as easily explain why

---

[52] Bach (1981) and Johnston (1988) both put forward non-intentionalist views on which they claim the agent *does* recognize the likely falsity of the proposition they desire to be true, but as we shall see, to bring self-deception to this level of awareness and reflectivity may point more strongly to a modified intentionalist account of self-deception, rather than a non-intentionalist one.

it appears that agents *in the grip of self-deception* characteristically exhibit cognitive tension.[53] As with Mele's account above, Barnes' self-deceiver is, by his own lights, perfectly rational. Given that the operation of the biasing mechanism in question occurs on the non-intentional level, the agent does not appear to be aware of its activity in producing his belief. If he were, he would likely take himself to have good ground to reject or revise his belief (or at least to view it more critically). Yet Barnes herself claims that, in cases like Jerry's, the self-deceived agent exhibits a kind of "false consciousness"—i.e., a failure to estimate highly enough to what degree his belief (or change in evidential standards) has been caused by the workings of his anxious desire on his reasoning processes. She asserts that the self-deceived agent may believe that his anxious desire plays no causal role in the formation of her belief, or that it plays less of an essential role than it actually does, or he may fail to have a belief either way as to whether his anxious desire plays a causal or an essential causal role.[54] Yet this "false consciousness" is not willfully or intentionally implemented. The agent is merely *ignorant* of the fact that her beliefs (or lack thereof) regarding her belief-formation are either mistaken or are themselves the result of a further act of biasing.

Yet self-deceivers appear to struggle cognitively in ways that merely biased believers do not. Scott-Kakures makes this point in his 2009 paper, "Unsettling Questions: Cognitive Dissonance in Self-Deception":

> A self-deceiver often, for example, displays remarkable credulity and resistance in her effort to settle a question of the form "p or not-p?" Evidence that strikes us as pathetically non-probative is frequently regarded by the self-deceiver as a sufficient basis upon which to settle her question. … Moreover, in other circumstances, self-deceivers resist the import of data, data that strike us as obviously sufficient for the settling of the question. Indeed, very often we can "rub the noses" of self-deceivers in what we take to

---

[53] Cf., for example, the passage by Shapiro I quoted in Chapter Two.

[54] Barnes, op. cit., 100-1.

> be the truth, only to provoke renewed and intensive investigations that certainly appear to be purposefully directed toward the embrace of falsity. … In many such cases, as well, the self-deceiver's efforts to settle her question are vexed, tedious, time-consuming, and so, aversive. What motivates her to persist in her struggles in the face of such persistent difficulties when the answer to her question is…right there in front of her eyes?[55]

These qualities of self-deceivers are not foreign to most of us, yet it is not clear how either Mele's or Barnes' deflationary accounts of self-deception can make sense of the psychic tension exhibited by all or most self-deceived agents while in the grip of self-deception.[56] Although we can imagine that, on a deflationary account of self-deception, an agent may come to suspect that she has acquired a belief in a motivationally biased manner, and that this discovery may cause her some cognitive discomfort, it appears that this can only happen *after* the self-deception is complete. And, as Scott-Kakures points out, the failure to make room for cognitive dissonance in self-deception itself "is to have lost what is most distinctive and vexing about self-deception."[57]

3.322 The Maintenance of Self-Deceptive Beliefs & Internal Irrationality

This seems to be a compelling reason to favor intentionalist over non-intentionalist accounts. However, let us grant for the sake of argument that the non-intentionalists are right about some cases of what we might call cases of "momentary" or "initial" self-deception.[58] We may suppose that Jerry's anxious desire (or net of desires, on Mele's account) non-deviantly causes him to treat the evidence in motivationally

---

[55] Scott-Kakures (2009), 74.

[56] Mele admits that deceiving oneself "might *often* involve considerable psychic tension," but he does not explain in what this tension is supposed to consist, nor how or why it would arise on his account. (Cf. Mele, 1997, 131).

[57] Scott-Kakures (1996), 37.

[58] As we shall see in Chapter Four, I prefer to classify instances of so-called "momentary" or "initial" self-deception as instances of "wishful thinking" or "hot biasing," precisely to distinguish them from what I take to be self-deception proper. However, the former terms will suffice for the moment.

biased ways, resulting in the belief that he does not desire Elaine and allowing him to maintain his self-image of being a good husband.[59] However, self-deception is rarely a one-time event. Because many irrational beliefs do not line up with the evidence the agent continually encounters, they are particularly unstable and need to be constantly fostered, reaffirmed, and reiterated. Thus, even if Jerry were non-intentionally to arrive at the false belief that he does not desire Elaine via a motivationally biased manipulation of the evidence, it is not clear that he could *maintain* this self-deceptive belief without somehow *strategically* remaining ignorant of either his desire for Elaine or of the fact that his belief that he does not desire her was irrationally caused. On an account like Barnes's, we must assume that, if Jerry's belief is maintained over a period of time, every time new evidence comes to light that he desires Elaine, he must be said to (non-intentionally) deceive himself anew (insofar as each new piece of negative evidence following a self-deceptive act may generate a new anxious desire, to which Jerry must respond).

Mele claims that explaining how a person retains a certain self-deceptive belief "need not be significantly different from the project of explaining his acquiring it, even if the belief persists for quite some time. …. The general phenomenon of remaining in a state of self-deception does not require an explanation that is different in kind from a proper explanation of entering self-deception."[60] But even if the psychological processes involved in maintaining a state of self-deception do not differ significantly from those involved in initially entering that state, this does not mean that these processes are

---

[59] As mentioned above, I do not wish to deny that phenomena like wishful thinking, motivational biasing, and so on ever occur. Surely many of our beliefs are not only motivated but also more-or-less directly caused by certain of our strong desires, fears, and/or other conative/emotional states. And if this represents a type self-deception, then non-intentionalist accounts do get at a certain kind of way of being self-deceived. But, again, the point of this section is to show that their view fails to plausibly account for certain (other) self-deceptive phenomena.

[60] Mele (2001), 46.

themselves sufficient to sustain the self-deceptive state. Indeed, it seems that in many cases *more* effort may be required to maintain a particular self-deceptive belief, especially if evidence continues to amass against the favored belief (assuming the agent is still responsive to evidence in general). Thus, in some cases, mere motivational biasing may no longer be enough to maintain a particular belief—especially if a certain degree of awareness of one's own irrationality creeps in.

Take instances in which non-consensual sexual relations between a parent and a child go "unnoticed" by other family members for long periods of time. Imagine the case of Rosie, who married Harry when her son, Oliver, was five years old.[61] Suppose also that Harry rescued Rosie and Oliver from a life on the streets and now provides them with all the material comforts of a middle-class family. However, within a year of the marriage, Harry begins to sexually molest Oliver. Oliver experiences both shame and guilt, and out of a fear of having to return to the streets, says nothing to his mother. Harry also says nothing to her. At first, Rosie may be completely ignorant of her husband's actions. She is initially a victim of interpersonal deceit[62] by both Harry and Oliver (who hide the incest for different reasons, of course). But as the evidence amasses over the course of several months, Rosie may find herself with an anxious desire, in the sense characterized by Barnes. Perhaps Harry's sexual interest in her has waned; he no longer sleeps in her bed; Oliver has become quiet and withdrawn, and she finds it difficult to explain the odd bruises and marks she finds on her son.[63] At this point, Rosie, who

---

[61] I have borrowed the scenario from an episode of "Law & Order: Special Victims Unit," but examples of self-deception among (both direct and indirect) victims of sexual abuse is not uncommon. See, for example, Nachson (2001) for a detailed discussion of deception and self-deception among victims and perpetrators of sexual abuse.

[62] This will likely be a kind of deceit by omission, though it could involve explicit lying.

[63] It is also likely that Rosie will be the victim of explicit interpersonal deception as Harry and Oliver must respectively explain their strange behavior. But this only goes to show the acute (and common) interplay between self- and other-deception in cases such as these.

desires to have a happy, normal family, may begin to fear that this is not really the case. She may also want to remain in the comfort of her current life, but fears having to return to the streets if something goes wrong with Harry. These (and other) anxious desires may causally contribute to her forming the false belief that Harry is a good father and that everything is fine with her son.

For the sake of argument, let us agree with Mele, Johnston, Barnes, et. al. that we need not suppose that Rosie is aware of what she is doing when she initially forms her self-deceptive belief against the evidence—at least not in a sense that would indicate a relevant intention to deceive herself. She merely treats the data in a biased manner, thereby causing herself to acquire a false belief that functions to alleviate her anxiety. However, when this false belief persists in the face of (often overwhelming) evidence to the contrary over months or even years, it becomes more and more difficult to suppose that Rosie really is as "in the dark" regarding the irrationality to which her belief commits her as the deflationist would have us suppose.

In such a scenario, Rosie's motivated belief-state is precarious and unstable. If she is at all rationally sensitive to the evidence, it is likely that she will continually encounter her belief not merely as outstripping her evidence but also as directly contradicting what she (in some sense) takes that evidence to entail. Thus, we would expect Rosie to exhibit significant cognitive tension, not only prior to the acquisition of her wishful belief, but also *as a result* of the acquisition of that belief. But if this is so, it appears that this kind of cognitive tension may, in at least some cases, be characteristic of self-deception itself, as opposed to merely motivating it. Indeed, it becomes difficult to understand how Rosie can remain fully ignorant of her own irrationality in such a case. It seems more plausible to attribute to her some level of active participation or purposive collusion in her own deception. It appears that she is not merely negligent due to ignorance; she is, minimally, *complicitly* (if not straightforwardly *willfully*) negligent. She engages in practices which violate epistemic standards that she herself holds, and it is plausible to attribute to her a

level of awareness (in the sense of a strong suspicion, well-grounded fear, or even belief that Harry is molesting Oliver) that is not so straightforwardly explained on the deflationary model. Thus, if the type of cognitive tension relevant to self-deception really does indicate a level of internal irrationality on the part of the agent (in the sense of an agent's knowingly or willingly violating her own epistemic norms), we may have reason to think that this is a genuine feature of self-deception. And if this is so, we appear to be able to shift the burden of explanation back to the non-intentionalists, who must then show how such internal irrationality is possible within the context of their deflationary theory.

Of course, it is always open to the non-intentionalist to deny that the kind of awareness discussed above implies any sort of intent on the part of the agent to undertake or persist in self-deception. Rosie may be a sort of helpless bystander, who cannot resist the force of her (anxious) desires. And if this were the case, we would expect her to experience significant levels of cognitive conflict. However, this way of characterizing her condition assimilates self-deception to some kind of psychological compulsion. If (as both Mele and Barnes insist), the self-deceived agent can resist her deception, then she must be capable of resisting the operation of whatever biasing mechanisms underwrite her belief. And such a capacity, in turn, appears to require that one be able to detect and either prevent or put a stop to the activity of such mechanisms—i.e., to treat the evidence in rationally-sensitive ways.

The proponents of the deflationary theory are thus in a bind: If Rosie is competent to detect the activity of certain relevant biasing mechanisms, so as to be able to put a stop to them, then why doesn't she do so? If she simply fails to notice the activity of these mechanisms, then she is merely ignorant, not irrational.[64] If, to avoid

---

[64] Barnes claims that "the biasing process not only produces the anxiety-reducing belief that *p*, it also…prevents the subject from recognizing the extent to which the self-deceptive belief that *p* is due to its tendency to reduce anxiety" (Barnes 1997, 123). Thus, the self-deceived agent

this conclusion, non-intentionalists appeal to further motivational biasing mechanisms, they appear to fall prey to an infinite regress, insofar as they will continually have to appeal to further biases to explain the operation of any given bias, and if they try to put an end to the regress by insisting that those like Rosie *allow* the mechanism to operate as it does, this appears to import a level of awareness and intentionality that is hard to reconcile with non-intentionalism.

3.33 The Reflective Nature of Self-Deception.

Another reason that "self-deception is not so straightforwardly explained" is that self-deception appears to hinge on an agent's ability to *believe and act for reasons*. Self-deceived agents do not just "happen" to be self-deceived. They, like other agents, have reasons that may contribute to (and help sustain) their self-deceptions. To see that this is so, we need merely remind ourselves of the significant role that *rationalization* plays in self-deception. Self-deceived agents search out reasons for their beliefs and actions—otherwise they cannot view these beliefs and actions as *theirs*. And this phenomenon points to another important feature of self-deceptive irrationality that non-intentionalist accounts like Mele's tend to ignore: the distinctively *reflective* nature of self-deception. As Scott-Kakures (2002) notes, deflationary accounts of self-deception do not appear to be able to distinguish between non-rational animals and rational agents.[65] Non-human animals may process information in motivationally biased ways that satisfy Mele's jointly sufficient conditions for self-deception, but we do not take such animals to be capable of self-deception.

To borrow an example from Scott-Kakures, suppose my dog, Pablo, mistakes my rattling a cereal box for my rattling of his dog food when he is particularly hungry (but at

---

appears to be characterized by a level of ignorance that may preclude her from resisting her irrational belief.

[65] Scott-Kakures (2002), 580.

no other time), and suppose that his hunger (or desire to eat) non-deviantly causes a false belief in him that I am rattling his food, so that Pablo may be said to process this information in a motivationally biased manner. Such a case appears to satisfy all four of Mele's jointly sufficient conditions for self-deception. But surely Pablo is not self-deceived. So what is missing from Mele's account? Scott-Kakures argues that deflationary theories of self-deception generally fail to take note of the role that reflective, critical reasoning plays in self-deception. Whereas agents *reflect on* and are *moved by* their reasons for acting and believing, non-rational actors "cannot be moved by reason *qua* reason, by the thought that *this* is a reason for believing *that*."[66] Thus, he argues**,** "one must understand what it is for one thing to be a reason for another"—to have "the sort of self-awareness necessary for thinking that one has good enough reason for believing something"—to count as self-deceived.[67]

Barnes is well aware of this fact. For this reason, she requires that in self-deception "one believes…that one's [self-deceptive] belief that *p* is justified."[68] This, of course, rules out that non-rational actors like Pablo can be self-deceivers, for Pablo is incapable of evaluating his beliefs. Barnes adds that "as soon as a child can understand that one thing can be a reason for another, the child can self-deceive him- or herself."[69] However, it is not clear what role this capacity for understanding reasons really plays in her account. In the end, it seems that, for Barnes, the only difference between Pablo and a self-deceiver is that the former's beliefs are caused more or less directly by the biasing mechanism in question, whereas the latter's are caused by a biased assessment of the

---

[66] Ibid., 585.

[67] Ibid., 581, 582.

[68] Barnes (1997), 117.

[69] Ibid., 117, n.11.

weight of the evidence. But again, Barnes makes self-deception look a lot more like a case of making a mistake than of actively manipulating one's reasons for believing.[70] So just as Pablo's desire causes the mistaken belief that his food is being rattled, the self-deceiver's anxious desire non-intentionally causes a mistaken belief regarding the nature of the evidence. But this represents no major difference in the *way* in which these false beliefs are produced, so to make a distinction between non-rational actors and self-deceived agents really amounts to nothing more than saying that the latter is potentially sensitive to evidence whereas the former is not.

Presumably, the primary reason Barnes wants to restrict self-deception to actors who can take one thing to be a reason for another (other than the fact that it strikes us as rather implausible that non-rational actors like Pablo should be self-deceived in the first place) is Barnes' intuition that self-deceived agents are, in fact, *responsible* for their self-deceptions. This can only be the case if we charge the self-deceiver with "epistemic negligence." Self-deceived agents fail to believe as we think they ought. According to Barnes, they *underestimate* the role that a particular attitude plays in their acquisition or maintenance of particular beliefs, and insofar as self-deceived agents are competent to avoid falling prey to this kind of error, they believe in ways that are "epistemically irresponsible,"[71] whereas Pablo is simply incapable of being responsible for his motivationally biased belief. This is an important step in the right direction for the deflationary theorist.

However, we often think that self-deceivers are more strongly responsible for their deceptions than a charge of mere "negligence" or "irresponsibility" indicates. Whereas cold biasers and wishful thinkers may simply not notice the activity of a

---

70 "The self-deceived *misapprehends* the structure of his attitudes," Ibid., 99. My emphasis.

71 Cf. Ibid., 83-7.

particular biasing mechanism at work in their belief-forming processes, self-deceivers appear to be actively engaged in their deceptions (especially in cases of belief maintenance) Barnes' herself claims that being epistemically responsible involves "taking appropriate care" to prevent oneself from being biased in the ways involved in self-deception or to detect these biases when they occur—where by "taking appropriate care," she means:

> trying to resist, given one's anxious desire, the skewing of the belief-acquisition process in favor of certain beliefs. If a self-deceiver is epistemically irresponsible, then he does not put up sufficient resistance on a particular occasion, or he has allowed himself to become the kind of person who finds it natural not to resist such skewing.[72]

Barnes goes on to write that whether or not we charge self-deceivers with being epistemically irresponsible "depends on such things as how hard they tried and why they failed."[73] But notice here the heavily intentional language involved in such a description of epistemic responsibility in the case of self-deception. The notions of 'trying' and 'allowing' appear to attribute at least a degree of intentionality to the agent in question. This certainly does *not* entail that self-deceivers who do not resist their deceptions do so intentionally, but it does indicate that Barnes might take there to be a level of reflectivity in self-deception that is not accounted for on Mele's model. If so, this would be a step in the right direction.

But there is still something a bit puzzling about this notion of "trying" to resist one's self-deception—especially when we pair it with the reflective nature of self-deception. To see why, let us examine a few possibilities by using the above example of

---

[72] Ibid., 83.

[73] Ibid. 87, n.26.

Jerry, who is self-deceived in believing that he does not desire his coworker, Elaine. Call Jerry's anxiety that he is not a good husband *q* and his potential self-deceptive belief that he does not desire Elaine *p*.[74] [75]

> *A1*. Jerry is not aware of his anxious desire *q*. Jerry's having *q* non-deviantly causes Jerry to treat the evidence in motivationally biased ways, resulting in his believing that *p*. He thus puts up no resistance to the belief that *p*, believing as he does that his acquisition of *p* is justified.
>
> *A2*. Jerry is aware of his anxious desire *q* but fails to realize that *q* may cause (or is causing) him to manipulate *p*-relevant data. He thus does not put up any resistance to the belief that *p*, believing as he does that his acquisition of *p* is justified, and he thus comes to believe that *p*.
>
> *A3*. Jerry is aware of his anxious desire *q* and, knowing that he has a tendency to bias evidence in favor of maintaining his self image, he tries to resist manipulating evidence regarding his desire for Elaine, but either the operation of *q* or the pull of *p* is simply too strong, and he fails, with the result that he believes *p* regardless.
>
> *A4*. Jerry is aware of his anxious desire *q* and, knowing that he has a tendency to bias evidence in favor of maintaining his self image, he tries to resist manipulating evidence regarding his desire for Elaine. Further, he succeeds in preventing himself from concluding that *p*. Instead, he forms the rational conclusion that not-*p*.

Let us look a bit more closely at *A1-A4* to see in which cases we might be inclined to claim with Barnes that Jerry is epistemically (ir)responsible. *A1* and *A2* both present us with cases in which Jerry is ignorant of his anxious desire and/or of its workings on his belief-formation processes. But we would not expect Jerry to resist concluding that *p* when such ignorance is in play, and we would not likely blame him for it unless we think his ignorance is itself motivated or the result of self-deception. But then Jerry is not really

---

[74] Or one may characterize the self-deceptive belief as a belief in the nature of the evidence, as we have done above.

[75] We can alter each example to reflect cases in which Jerry has already self-deceptively come to the false belief that *p*, where the issue at hand is whether or not Jerry fails to detect his irrationality and thereby to rationally revise his belief.

epistemically irresponsible in failing to resist the belief that *p*; Rather, he is irresponsible insofar as he is ignorant in one or both of the above ways. In such a case, we might, of course, blame Jerry for having "allowed" himself to become the *kind* of person who fails to detect anxious desires or their tendency to bias him in non-truth-conducive ways, or who "finds it natural not to resist such skewing."[76] But this kind of "allowing", too, appears to be an active notion that requires a level of awareness, for if Jerry really could not help becoming this kind of person, we might not be inclined to hold him epistemically responsible for these character traits either. Thus, a charge of epistemic negligence or irresponsibility in cases of the type represented in *A1* and *A2* appears to require a level of active participation or complicity on the part of the agent—and this, in turn, appears to require a level of awareness of what one is doing (or of the kind of person one is becoming). But this level of reflective awareness appears to be missing on Barnes' and Mele's accounts.

What about the cases of *A3* and *A4*? In both cases, it is supposed that Jerry is aware both of his anxious desire, *q*, and of its potential (or actual) effect on his belief-forming processes regarding *p*. Furthermore, in both cases, Jerry tries to resist the effects of *q* on his beliefs. However, in *A3*, Jerry is simply too weak to resist the biasing effects of *q*, either because his anxious desire is too strong to resist, or because *p* represents an overwhelmingly attractive way of assuaging the anxiety wrought by *q*. In this case, Jerry is cognizantly irrational, but it appears that he believes compulsively. But then Jerry is *not* competent to avoid his irrationality, which supposedly violates a condition of both Barnes' and Mele's accounts of self-deception.[77]

---

[76] Ibid., 83.

[77] Barnes is somewhat unclear on this point. In a note to Chapter Three of *Seeing Through Self-Deception*, she claims that "self-deceptive belief is always other than a compulsive belief" (46). However, in Chapter Five, she maintains that "if an anxious desire is powerful enough, it seems possible that it will so bias the person's thinking that all the evidence will only further strengthen the person's belief that his thinking is not biased. In such circumstances, given

Finally, *A4* appears to be an instance in which Jerry exhibits a kind of praiseworthy (or at least not blameworthy) epistemic responsibility. However, notice that here Jerry is cognizant of both his anxious desire and the ways in which that anxious desire may skew his evaluation of the evidence. He is "on the look-out" for the biasing he knows may accompany the presence of such recalcitrant desires, and he takes pains to correct the operations of his anxiety on his belief-forming processes. And thus *A4*, as opposed to *A1-A3*, does not represent an instance of self-deception, on the deflationary picture, insofar as Jerry comes to the rational belief that not-*p* by exercising his "competence" to resist the self-deceptive belief that *p*.

Of course, there are other possible scenarios. Perhaps Jerry just happens to successfully resist believing that *p*, even though he is, e.g., ignorant of his tendency to manipulate data in the presence of an anxious desire, or of the fact that he has the anxious desire that *q*, etc. But surely this kind of "accidental" resistance is not what Barnes has in mind. Yet there is another scenario we have not considered. Perhaps Jerry *is* cognizant of his anxious desire, *q*, and the biasing effect that *q* is exercising on his beliefs regarding the evidence. However, Jerry does *not* try to resist the biasing effects of *q* and instead

---

the power of the anxious desire involved, it is not clear that there is anything that the person could have done that would have prevented the biasing processes from being successful…[I]t is not clear that the person in such a situation would be epistemically irresponsible" (87).

I take it that the difference between the Chapter Three and Chapter Five claims rests on the respective difference between a case of the type described in *A4* and a stronger version of the types described in *A1-A2*, in which the agent does not realize her thinking is biased, due to the strength of her anxious desire. Of course, what is important here is that if Barnes is willing to allow Chapter-5-type cases to count as instances of self-deception, then it is not at all clear that the agent must be competent to avoid becoming self-deceived (or to avoid becoming the kind of person who is so strongly biased) to count as self-deceived. But then it is not clear what work Barnes' condition that the self-deceiver believe her self-deceptive to be justified is doing in this context. Presumably, she needs this condition to explain what makes self-deception irrational, as opposed to merely *a*rational—or she is merely appealing to the phenomenology and observed behavior of self-deceived agents. But the presence within her theory of cases of self-deception which the agent is incompetent to avoid make some instances of self-deception look much more akin to Pablo-type cases, at least as far as the *causal* story is concerned.

allows them to operate as they do. Here, of course, Barnes would say that Jerry is epistemically irresponsible. However, assuming he acts *freely* (that is, he has it in his power to resist manipulating the evidence in a motivationally biased manner—i.e., to act/believe *rationally*), the static and dynamic puzzles again read their ugly heads, for now it appears that Jerry knowingly and willingly allows himself to violate his own epistemic standards—that is, he tries to believe for reasons he himself takes to be inadequate reasons for believing!

Yet it is precisely this kind of phenomenon that appears to distinguish the self-deceiver from the merely ignorant or compelled agent (or from the non-rational animal). It is not *merely* the fact that the former is both motivated and in some sense "competent" to avoid her deception, but also that in self-deception the agent actively *puts reason to use against itself* in ways that ignorant, compelled, or non-rational actors do not (and in the latter cases *could* not). The self-deceived agent "gives himself reason for believing what he believes he has sufficient reason not to believe."[78] This is also a way of making the distinction between mere wishful thinking and self-deception proper. Wishful thinking does not require the presence of reflective, critical reasoning, whereas self-deception most certainly does. In wishful thinking, we may think of reason as being "hijacked" by one's desire, whereas in self-deception the agent is a "willing participant" in the acquisition or maintenance of some favored belief.[79] Thus, the distinction between wishful thinking and self-deception can be viewed as a difference in *kind*, not merely in *degree*, as Mele would have it. Furthermore, making this distinction can explain why self-deceived agents exhibit cognitive tension of the kind mentioned above. Insofar as the self-deceiver is employing distinctly rational (i.e., reasoning-giving) techniques in the

---

[78] Scott-Kakures (1996), 591.

[79] Cf. Ibid., 585.

service of *ir*rationality, it is no surprise that she would experience being "at odds" with herself—for she is, in fact, undermining her own rationality by employing the very faculties that make her a rational agent in the first place. She is, as we have postulated, *internally* irrational—that is, she is irrational *by her own lights*.[80]

It should come as no surprise that the process of self-deception should put certain rational faculties to use. Indeed, what is so puzzling about supposedly paradigm cases of irrationality like self-deception and weakness of will is that they appear to rest on or represent actions performed for a reason, yet from the agent's own point of view these reasons appears inadequate or insufficient to rationally ground or justify the action performed. But the intuition that non-rational actors cannot act irrationally points to the notion that reason can somehow work to undermine itself in some cases. I shall have more to say on this in the next chapter. What is important to note here is that however one cashes out the origin, purpose, or function of our capacity for self-deception, it does not seem to be independent of our capacity for rational, reflective, critical thought.[81] And any adequate account of self-deception will have to take this into account.

3.4 Conclusion

To sum up what we have said thus far, the claim of proponents of the deflationary account that their account is explanatorily more parsimonious and therefore preferable appears dubious. There appear to be phenomena that these accounts do not adequately address, including (though not necessarily restricted to): the cognitive tension apparent in the behavior of self-deceived agents, the distinction between ignorant or compelled agents and self-deceived agents, the intuitive difference in kind between wishful thinking

---

[80] For more on the importance of internal irrationality to intentionalist accounts of self-deception, see Scott-Kakures (1996).

[81] Neil Van Leeuwen (2007) argues, for example, that the capacity to deceive ourselves is a kind of "evolutionary spandrel" that developed as a biological "offshoot" of the neurological capacity for rational thought.

and self-deception proper, and the reflective nature of self-deception. For this reason, one might conclude that a more robust account of self-deception is necessary to make sense of these phenomena. It is to such an account that I now turn.

CHAPTER 4. THE INTENTIONAL PROJECT: TOWARD A DIFFERENT UNDERSTANDING OF SELF-DECEPTION

4.1 Some Important Intuitions Regarding Self-Deception.

Thus far, we have examined two prominent approaches to self-deception and have found both to be inadequate. In Chapter Two, we saw that, in their attempts to make room for the intuition that self-deception is both intentional and internally irrational, partitioned-mind theories raise worrisome metaphysical problems regarding the nature of the self, agency, and action. And in Chapter Three, we saw that non-intentionalist theories regarding self-deception escape the straightforward paradoxes of irrationality only because they fail to account for certain crucial features of self-deceptive irrationality.

Even if we reject both accounts, however, we should not overlook the very important motivations that underlie them. The non-intentionalists note, first, the conceptual and psychological difficulty involved in intentionally acquiring beliefs, especially beliefs one takes to be unwarranted by the evidence. Second, they point to the empirically demonstrable fact that human beings very often believe against the evidence and, as such, are much less "rational" than philosophers have traditionally taken them to be. (Of course, the kind of irrationality that concerns them is the "external" or "weak" irrationality of holding beliefs that are objectively unwarranted by the evidence, or beliefs that would be considered unwarranted by the vast majority of one's impartial cognitive peers.) Partitioned-mind theorists, on the other hand, characterize self-deception as representing a paradigm case of intentional, *internal* irrationality, in which the agent is, in a very literal sense, "divided" against herself. As we have noted in previous chapters, self-deceivers typically exhibit a certain kind of cognitive dissonance or psychic tension

that might indicate the presence of a stronger kind of irrationality than that pointed to by the non-intentionalist. A related advantage of, intentionalist strands of partitioned-mind theories is that they preserve the analogy between self- and other-deception.

It is important to note here that none of these driving intuitions is, strictly speaking, incompatible with the others. Non-intentionalists simply point out that doxastic voluntarism (in the sense of *directly* causing oneself to have a particular belief) is generally false[1] and that human beings often believe against the evidence, whereas partitioned-mind theorists claim that the existence of a stronger kind of irrationality is needed to make sense of the phenomena and that self-deception may, in fact, mirror interpersonal deception in some way. The question is whether we can arrive at an understanding of self-deception that accommodates all these intuitions without running into the difficulties presented by the static and dynamic paradoxes as we described them above. In other words, is there a way to incorporate the best elements of both types of theories we have examined, such that we arrive at an account of self-deception which points not only to something that is psychologically possible but also to something we think is psychologically *plausible*? I think there is a way, and in what follows I will try to sketch just such a view.

4.2 Toward a Diachronic Account of Self-Deception.

One of the biggest problems with the prevalent literature on self-deception is that it tends to ignore the *process* by which an agent may be said to be self-deceived. To be

[1] I do not wish to contest this claim here. In general, I think the intuition that doxastic voluntarism is false is correct. While there may be situations in which something like "willing to believe" might be psychologically possible, I think these cases will be few and far between—and thus cannot account for the widespread prevalence of self-deception among human agents. For conflicting views on the possibility of willing to believe, cf. James (1979) and Williams (1973).

sure, non-intentionalist and intentionalist accounts both have some story about how agents may arrive at or retain irrational beliefs, but these strategies and processes are often viewed as *leading up to* self-deception, not as being *constitutive of* self-deception itself. I think this is a mistake. Indeed, if we move from approaching self-deception as a kind of static "condition" or "state" toward viewing it instead as a kind of diachronic "project" or "undertaking" in which an agent actively engages, the supposed problems with self-deception may dissolve altogether.

Let us reexamine the example of Agnes and Ralph from Chapter Two to see how this might work. Recall that Agnes, who has heretofore never taken herself to have evidence that her husband, Ralph, is being unfaithful to her, has recently encountered strong evidence to the contrary (e.g., he smells like another woman's perfume, she has found lipstick stains on his collar, he makes implausible excuses for his unusual behavior, she has been informed of his dining with another woman, etc.). Agnes now has ample evidence to suspect Ralph of cheating on her, and, indeed, she does begin to question his fidelity. That is, her assessment of the evidence points toward his having an affair. Agnes' belief that Ralph is faithful stands on shaky ground, and it is clear she should consider rejecting it, holding instead that he is unfaithful. However, Agnes also has a strong (and quite understandable) desire that he *not* be cheating on her. She values her role as a beloved wife, and her husband's love plays an important part in her ability to maintain such an image of herself. It is here that the project of self-deception may get set up. Instead of revising her belief, she may attend more closely to fulfilling her desire that he not be cheating on her. Of course, she recognizes the impossibility of doing this directly, given what she takes the evidence to demonstrate, and given the difficulty of

believing a particular proposition at will. Therefore, she must take steps to *indirectly* cause herself to believe that Ralph is faithful.

A non-intentionalist like Mele might claim that Agnes may, in fact, *intentionally* avoid evidence that Ralph is cheating because, say, she finds it discomforting, and he might note that this could explain why she retains her belief in Ralph's fidelity . In short, there is no need to assume that she needs to intend *to retain this belief*. However, with every new piece of evidence that crops up against Ralph's fidelity, Agnes will have to deal with her problem anew. And if the evidence continues to point in this direction while Agnes continually fails to believe it, it does not seem prima facie implausible to attribute to her a kind of long-term intention to believe in Ralph's fidelity, come what may. In other words, it is very likely that, at various points in her attempt to believe in her husband's faithfulness, Agnes recognizes (or at least suspects) that the evidence does not favor her current belief, and she also likely recognizes that she has no control over this fact. But she *does* have indirect control over what she believes, insofar as she can (to some extent) control what she takes to be reasons for her belief by selectively directing her attention, rationalizing, engaging in positive thinking, and so on. Thus, although she cannot easily change the world to fit her desire, she can take the means to try to bring it about that she *believes* the world is as she desires it to be.[2]

---

[2] Of course, Agnes could lock Ralph in his room or even kill him, thereby ensuring he does not cheat on her in the future, but presumably this is not an attractive or viable alternative for her, given her other beliefs and attitudes. Additionally, for Agnes, what falls under the proposition "Ralph is faithful" may be something like that Ralph does not *willingly* cheat on her when free to do so, or that he has never cheated on her in the past. And given her inability to control Ralph's free actions or to change the past, attempting to change the way she views the world may appear the more attractive course of action.

The project of deceiving oneself differs from other projects in one important respect: in order to deceive oneself one must violate one's own epistemic norms.[3] This does not disqualify it from being an intentional project, however. My point, moreover, is that deceiving oneself just *is* pursuing such a project.

4.3 Dealing with the Paradoxes.

Does such an account of self-deception escape the static and dynamic paradoxes we mentioned above? I think it does. First, there is no requirement on this account that the self-deceiver believe contradictory propositions. The self-deceiver may not have a full-blown *belief* that the evidence supports not-*p*; rather, she fears that it might, and rather than pursue this possibility, she makes a concerted effort to establish that *p*. Returning to Agnes, we may suppose that she begins with the belief that Ralph is faithful to her. As the evidence mounts up against this belief, Agnes comes to suspect that he might be cheating. Her belief in his fidelity is weakened. How might she go about continuing to believe her husband is faithful? She might rationalize away the evidence or focus on other evidence that supports her belief. Her assessment of the evidence leads her to *fear* that not-*p* is the case, and it is this fear that provides her with a reason to pursue a self-deceptive project may last for days, months, or even years. But nowhere along the line do we need to suppose she simultaneously holds contradictory beliefs. Indeed, it is likely her intentional attempt to *avoid* holding inconsistent beliefs that causes the cognitive tension so typical of self-deceivers.[4]

---

[3] I discuss the (ir)rationality of self-deception below.

[4] The reader may notice here a similarity to Barnes' account of self-deception as motivated by an "anxious desire," as discussed in Chapter Three. However, my account differs from Barnes, insofar as I claim the self-deceiver must have some level of reflective awareness regarding her assessment of the evidence, for as I argue in Chapter Three, if she were completely unaware of the role her evidential assessment plays in motivating her self-deception, she would likely not experience any significant cognitive tension in carrying out her self-deceptive project.

But what of the dynamic paradox? It appears that Agnes is in some sense aware of what she is up to in a way the wishful thinker is not. So how could she ever succeed in convincing herself that Ralph is truly faithful to her? In considering this challenge, it is important to note, first, that human beings are creatures of habit. The more a self-deceiver engages in her self-deceptive behavior, the more such techniques become a matter of habit. Take as an analogy learning to drive a stick shift. Initially, pressing down on and releasing the clutch so as to shift gears may take a lot of conscious effort—effort that may interfere with one's successfully driving down the street. However, the more one drives a manual transmission, the more routine and less strenuous such activities become. So it is with self-deception: The more Agnes becomes entrenched in her self-deceptive project, the less difficult it becomes to focus her attention away from disturbing evidence.[5]

Of course, a theorist of the non-intentionalist persuasion might object that Agnes's acquisition or successful maintenance of the irrational belief that Ralph is faithful is not itself, intentional, but is (as the non-intentionalist claims) the product of nonintentional motivated biasing (in this case, of a kind of wishful thinking). Strictly

---

[5] Of course, should Agnes eventually reach the point where *nothing* (barring some momentous event—e.g., walking in on Ralph and his mistress in the act of lovemaking) counts as evidence against her belief in his fidelity, I would argue that she is no longer engaged in self-deception, but is rather *delusional*. She believes against evidence that any normal, rational agent would take to point to the contrary proposition.

In such a case, Agnes might be said to be in a similar epistemic situation to the epistemically ignorant agent. Her belief is informed by what she currently takes to be good reasons, and she feels no cognitive tension in believing against what (by her earlier lights) counted as evidence against her belief. However, unlike the epistemically ignorant agent, Agnes may be culpable for her delusional belief, insofar as she reached this belief by epistemically irrational means. Additionally, given her habituation via engagement in self-deception, Delusional Agnes will probably be less likely to revise her belief in light of new evidence, whereas the ignorant agent, if rational, would likely do so. This might lead us to better compare Delusional Agnes to the psychologically compelled agent, who cannot (now) revise her belief, given that her self-deceptive techniques are so deeply entrenched. However, if her delusional belief is the terminus of an intentional self-deceptive project, we may still maintain that she is in some sense responsible (or at least answerable) for her having acquired this irrational belief. I return to questions of epistemic and moral responsibility for one's self-deception in Chapter Five.

speaking, this is correct.[6] However, this is again to focus on the terminal *state* of self-deception rather than on the entirety of the *process* by which Agnes reached her self-deceptive belief, and it is this latter activity that I have characterized as intentional, not the resulting mechanism by means of which she forms and/or sustains her belief. The self-deceiver is concerned with forming or maintaining the relevant self-deceptive belief in whatever way possible (regardless of the force of the evidence), and it is this that can be characterized as intentional. Thus, some deceivers may actually *intend* to go in for wishful thinking. That is, the terminus of an agent's intentional self-deceptive projects may be a belief acquired non-intentionally via wishful thinking or some other bias, but that does not make her self-deception itself unintentional, as it may be precisely this kind of belief formation at which the agent aims, given that she cannot directly will herself to believe the favored proposition.[7]

4.4 Engaging in Self-Deception vs. Being Self-Deceived.

The above discussion raises an important point, namely that on my account self-deception is not necessarily a reflexive self-relation. That is, there may be a distinction to be made between *engaging* in self-deception and *being* self-deceived. Actually, the term 'self-deceived' is ambiguous. In one sense, we employ the term to designate any person currently engaged in a project of self-deception. On this meaning, to say a person is self-deceived amounts to nothing more than saying that the person is engaging in self-deception. Of course, in such cases, we more often use the present progressive, e.g.,

---

[6] I will leave open whether the end product of the successful self-deceptive project will be a belief produced by a habituated biasing mechanism (e.g., wishful thinking), or a revaluation of evidence, which leads the agent to believe for what she now takes to be good reasons (assuming there is a relevant distinction to be made here). In either case, Agnes does not, in the end, appear to exhibit the internally irrationality we took to be characteristic of self-deception.

[7] As I discuss below, the last step toward the terminus of many of our intentional projects may not be directly preceded by an explicit proximal intention for the project to count as intentional. The success of many of our day-to-day intentional projects hinges on external, purely contingent factors over which we have little to no control.

"Steve is deceiving himself." But even in ordinary parlance, we sometimes denote agents who habitually engage in such projects as being self-deceived. In these instances, there is no significant distinction between engaging in self-deception and being self-deceived.

On the other hand, 'self-deceived' may be taken to mean having succeeded in one's project of self-deception (i.e., having successfully acquired or maintained the belief in question), and here there is a conceptual distinction to be made between deceiving oneself and being self-deceived. This is an important respect in which, on my account, deceiving oneself closely resembles deceiving others, and it is precisely this feature of the account that allows it to escape the paradoxes. Just as in interpersonal deception, an agent (*A*) may be said to engage in the deception of another person (*B*) without *B* thereby being said to be (or having been) deceived by *A*, self-deceivers may undertake intentional projects of self-deception without thereby becoming "self-deceived."

Not only is it the case that both types of projects can fail to bring about the intended result, but they both involve the same sort of activity, namely *persuasion*. In particular, like cases of self-deception, most cases of interpersonal deception require the deceiver to do things to make the relevant falsehood seem believable. That is, *A* must usually *persuade B* of the truth of the proposition (*p*) which A takes to be false. He must, e.g., act as if *p* is true, offer reasons for believing *p* which *B* will accept, and so on. And although we might hesitate to say that *A* actually *deceived B* regarding *p* in cases where the former was not able to convince the latter, we would not thereby absolve *A* of having engaged in deception. The same can be said of self-deceivers.

A further parallel between self- and other-deception is also worth noting here. *A* may be said to engage in the deception of *B* regarding *p*, regardless of the truth value of *p*. What makes *A*'s activity deceptive is that she *takes p* to be false and yet tries to persuade *B* of the truth of *p*. And we may likewise say that *B* has been successfully "duped" or deceived by *A* when the latter is convinced by the former, even if *p* turns out to be true. In a case like this, *B* will not be deceived regarding *p*, in the sense of being

mistaken about *p*, but he will have been deceived by *A*, and in this sense we may still call him deceived. The same is true of self-deception. An agent may deceive herself regarding a proposition that turns out to be true, and if she succeeds, she may still count as self-deceived, even if she now possesses an (accidentally) true belief. This departs from both Mele's and Barnes' accounts of self-deception, insofar as on both theories, to count as self-deceived, the relevant belief must be false.[8] But to accept this latter claim would be to severely restrict the class of persons that count as self-deceived. It seems more appropriate (and commonsense) to say that an agent may be self-deceived regarding a true proposition, insofar as she arrived at this state via a successful project of self-deception.

Of course, it may be difficult to say when an individual has been successful in her endeavors to deceive herself (i.e., is self-deceived in this second sense), as an agent entrenched in a project of self-deception is constantly engaged in a battle between holding the favored belief and manipulating evidence to the contrary. Suppose the agent is able to maintain the belief in question at time t_1 (e.g., by rationalizing away some threatening negative evidence), yet at time t_{1+n} must grapple with the problem anew. Does she count as self-deceived in the second sense at t_1? Has she really been successful? That is, is it the case that her self-deceptive project terminates in her being self-deceived at t_n, such that her renewed effort to deceive herself at t_{1+n} represents a new self-deceptive project? Or is her project a continued one—one which endures from t_1 to t_{1+n} (and beyond)? These are difficult questions, which arise not only in cases of self-deception but in cases of intentional action in general. Many of our intentional projects involve concrete ends, e.g., losing 10 pounds—and in these instances, it is reasonably clear when one has been successful. However, other projects may be aimed at the continued *maintenance* of

---

[8] Cf. Mele (2001), 50; Barnes (1997), 118.

particular states, e.g., being healthy. Here, although my goal is being healthy, my project may be said to continue, even when I have achieved this state, since the project aims at *remaining* healthy. And, as we have seen, projects of self-deception often resemble projects of this latter type. Thus, one may also count as 'self-deceived' in the second sense while continuing to engage in self-deception, but one need not. Not all self-deceivers are (or will be) self-deceived in the second sense discussed above, but those who are self-deceived will have engaged in self-deception.

4.5 The Intentional Component of Self-Deception.

Since there is considerable resistance to the claim that self-deception is an intentional activity, I want to take some time to discuss the "intending" in more detail. Once we understand what is involved in intending to do something, we will see that there is no reason to object to the thesis that self-deception is an intentional project. This is true, I believe, on any plausible conception of intention. I will, however, limit myself to sketching at least one way in which we may plausibly understand intentions and intentional actions. In so doing, I hope to strengthen the case for my account, while sharpening my critique of non-intentionalist conceptions of self-deception.

With Kant, I take the will to be the activity of reason in its *practical* capacity, such that intending to φ amounts to a matter of resolving (positively) the question of whether or not to φ, where the activity of resolving this question involves responding to the various reasons one takes oneself to have for and against φ-ing.[9] That is, an intending to φ involves a kind of practical *commitment* to φ-ing. Following Anscombe, I maintain that intentions themselves are neither beliefs nor judgments about what it is best to do, nor even about what one will do. The result of a particular piece of practical reasoning is thus not to be confused with the result of a particular piece of speculative reasoning.

---

[9] Cf. Hieronymi (2009) for a similar view.

Whereas beliefs and judgments settle the speculative question of whether or not *p*, intentions settle the very different practical question of whether or not to φ.[10] This is not to say that theoretical considerations play no role in practical reasoning. Agents commonly have beliefs about what it is good or desirable to do and these beliefs usually play a decisive role in how the agents settle the question of how to act. An agent may settle the question of whether or not to φ without extensively reflecting on her judgments regarding the worthiness of φ-ing. Of course, although an intention need not be the product of explicit, self-conscious deliberation, an agent must, in some sense, "see" the facts as presenting her with a reason to act. For example, I may decide to take a drink of water without any explicit deliberation regarding my beliefs about my current state of thirst or my judgments regarding the practical or moral worth of staying hydrated. And this deciding may still represent a reasons-responsive settling of the question of whether or not to take a drink of water—i.e., I can still drink the water for reasons that, if prompted, I would endorse as *my* reasons for doing so. Thus, we may view an intention as a type of reasons-responsive commitment to a certain course of action and intentional

---

[10] Ibid., 206. One might want to contrast theoretical knowledge here with what Anscombe calls "practical knowledge," which she supposes to be non-observational and to serve as "the cause of what it understands" (cf. Anscombe, 1957, 87-8). While I will refrain from weighing in on the observational/non-observational debate, it does seem that intending presents us with a different way of viewing our interaction with the world than that of purely speculative knowledge. Whereas speculative knowledge derives from the objects known, practical knowledge (as embodied in intentions) serves as the (at least partial) *ground* of what is known. This leads to another important distinction. As Richard Moran (2004) writes:

> [I]t is important to…this idea [of practical knowledge as the cause of what it understands] that the interpretation of 'cause' here in terms of *efficient* causes is at best a partial and misleading understanding of the sense in which one's intention to pick up some milk can be "the cause of what it understands". The point is not that the knowledge embedded in my intention helps to *produce* the movements that lead to the picking up of some milk, but rather that those movements would not count as my picking up some milk (intentionally) unless my practical understanding conceived of them in those terms. (47)

I return to the distinction between appeals to efficient and final causation in explanations of action in the section below.

action as action performed by an agent which she herself endorses or takes herself to have reason to perform.[11] But what is it to do something because you take yourself to have a reason for doing it?

Anscombe writes that intentional actions "are the actions to which a certain sense of the question 'why?' is given application."[12] Note here that the type of *why*-question under discussion here is not the *why* of pure efficient or material causation. When we ask an agent why she is acting as she is—or why she acted as she did, or why she proposes to act in a certain way in the future—we are not generally interested in the proximal causal mechanisms at work in the direct *production* of the action (or, in the case of proposed future actions, the intention) in question. We do not usually mean to inquire into what physical or psychological states resulted in her acting as she does, as we might when we ask why, e.g., one has sneezed. No, when we pose the question *why?* to agents regarding their actions, we are looking for a kind of *final* cause of the action.[13] We want to know what the agent is aiming at, or hoping to achieve, or trying to bring about in so acting.

---

[11] Note: I am not interested here in exploring the metaphysics of reasons in detail. Rather, as I go on to discuss, I am concerned with what is involved in taking oneself to have a reason to act.

[12] Anscombe (1957) 9.

[13] This is not to say that we never inquire into the efficient causes of actions. Scientists may be interested in the efficient causes of particular types of actions. Likewise, we sometimes wish to know what actually ended up producing the action in question. This is especially true when the voluntariness itself of the action is under discussion—i.e., when we wonder whether the agent really did act for a reason. In such cases, one might ask what the most relevant cause of the action was: a disease? a brain seizure? an uncontrollable psychological impulse? or an actual volition?

Indeed, I think it is precisely the distinction between efficient and final causation that leads (perhaps unnecessarily) to some of the tension between the non-intentionalist and intentionalist camps in the literature on self-deception. Whereas non-intentionalists are predominantly concerned with the efficient or productive causes of self-deception (e.g., the activities of motivating desires or biasing mechanisms), intentionalists tend to focus more closely on the teleological nature of self-deception.

As we have seen, many non-intentionalists concede that self-deception is goal-directed in some sense. Most non-intentionalists claim that self-deception is, indeed, purposive, just not intentional. They claim that the biasing mechanism or other productive cause of an agent's self-deception serves some (generally useful) psychological or biological function or purpose.[14] Annette Barnes, for instance, proposes (quite plausibly) that self-deception functions to reduce anxiety.[15] "A self-deceptive belief," she writes, "is an effect whose purpose is to alter its cause."[16] Barnes does not analyze the concepts of 'function' or 'purpose' in any great detail, maintaining only that self-deception is purposive insofar as "something (having a self-deceptive belief) which has a certain effect (reducing anxiety) is explained by the fact that it has that effect."[17] However, although this is a type of teleological answer to the question *why?*, it is not the kind of answer we tend to look for when inquiring into the actions of agents. Rather, we generally mean to be looking for the reasons the agent herself endorse. What matters is why she thinks that it makes sense to act as she does. Of course, functional reasons and agents' reasons can (and often do) coincide, but there is still an important conceptual distinction at work here. The former type of reason may present us with justifying reasons that are independent of what the agent takes herself to be doing, but we are concerned with the justification that the agent herself endorses for her acting as she does.

---

[14] For an opposing view, cf. Van Leeuwen (2008), who argues that the capacity for self-deception developed as a mere evolutionary byproduct or "spandrel" of certain adaptive capacities for rational thought, and thus does not have an additional adaptive function of its own.

[15] Cf. Chapter Four of Barnes (1997).

[16] Ibid., 60.

[17] Ibid., 60.

The latter type of justification answers the so-called *why? of rationalization*.[18] Note that this latter type of *why*-question is (at least theoretically) addressed to the agent herself and inquires into *her* justification for what she does. That is, the *why?* of rationalization does not merely inquire into reasons that *would* hypothetically justify the agent's action, but rather into the reasons she *actually* endorses (or would endorse if prompted).[19]

We need to ask ourselves, then: Does the *why?* of rationalization apply to cases of self-deception? Is self-deception the kind of thing to which this particular sense of the question *why?* "is given application"? The notion that in the example above Agnes believes *in order to* x (where *x* gives an answer to the latter type of *why*-question) initially appears odd. From the point of view of the Anscombian interrogator, we normally do not pose this type of *why*-question regarding the acquisition or retention of beliefs. As Hieronymi (2008) writes:

> In believing, you are answerable for reasons that you take to show the belief true. In contrast, in either intending to act or acting intentionally (where the action may be as complex as you like), you are answerable for reasons that you take to show something good about so acting. If we were to try to make belief into an action, one would have to be, in believing, answerable for reasons that one takes to show something good about believing. But these are a different class of reasons than those one takes to show that *p*. Because believing thus entails its own distinctive form of answerability, believing *p* cannot be understood as an action in its own right. To put the point somewhat differently, because

---

[18] That is, one may give the functional explanation that a particular action is to be explained because, e.g., it is conducive to the survival of the species. But this reason is not necessarily a justifying reason, unless the agent herself adopts and endorses this consideration as a reason in favor of so acting.

[19] For example, a functional explanation for the fact that two men engaged in a bar fight might be that they were genetically predisposed to "impress" or "compete for" the females in the room because, historically, this trait was conducive to the propagation of the human race. However, it is unlikely that they would endorse this fact as their reason for engaging in the bar fight.

> believing brings with it its own distinctive form of answerability, believing is not the proper object of an intention.[20]

So, as Hieronymi aptly points out, we are answerable for our beliefs in a different way than we are for our actions. We may ask for an agent's *epistemic* reasons for believing as she does, but we do not usually think that belief is the sort of thing that is voluntary, so we do not generally ask for an agent's prudential reasons for her belief. Thus, to ask why Agnes believes Ralph is faithful does not seem at first glance to be susceptible to the kind of reasons-answer we typically give for intentional actions.

Of course, when an agent's epistemic reasons fail to support the belief she holds, we may consider whether she has other types of reasons. If, for example, we believe that Agnes does not have good epistemic reasons for her belief that Ralph is faithful, we may reasonably wonder whether she believes he is faithful because she values her marriage, or because she wants to be loved, or because she fears Ralph will leave her, or because she wants to preserve her self-image. Such explanations do not imply that Agnes is aware of forming her belief on such grounds. Indeed (as the dynamic paradox has shown us) if she were to propose one of the above reasons as an answer to the Anscombian *why*-question, or were to adopt one of these prudential reasons as *her* reason for believing, this would threaten to completely undermine her project to so believe. In other words, she cannot give this reason as *her* justifying reason for believing as she does without endangering the very belief that this reason is supposed to pragmatically justify.

The intentional-project account of self-deception thus appears to be burdened with two closely-related worries:[21] First, belief is not generally voluntary, so the types of

---

[20] Hieronymi (2008), 355-6.

prudential reasons offered in explaining intentional action do not appear to apply to cases of believing, even self-deceptive believing. Second, even if the agent has non-epistemic reasons for believing as she does, those reasons do not appear capable of being *her* reasons (in the sense of being reasons she takes herself to have) Does it follow that self-deception cannot be an intentional activity, after all? To answer this question, we need to take a closer look at intentional actions —especially intentional actions of longer duration.

First, although it may be true that one cannot straightforwardly acquire beliefs voluntarily, this does not mean that one cannot initiate an intentional project *aimed at* coming to or continuing to believe. Although it might not be possible to acquire a particular belief via a single act of will (or via what Mele and Moser call a "proximal intention," an intention for the specious present[22]), this does not mean that one could not initiate an intentional project that results in one's holding a certain belief. As Hieronymi points out, although you may not be able to have a direct intention to believe that *p*, you *can* have an intention "*to bring it about* that you believe *p*, an intention to make it the case that you settle the question of whether *p* positively, and so believe. Bringing something about is an ordinary, voluntary action [even though] believing [itself] cannot be."[23] And we might apply the general point to the specific case of self-deception: a self-deceptive project is one in which the agent aims to *bring about or maintain* a belief she

---

[21] It might turn out that these are really just different aspects of the same problem. However, even if this is right, it may be facilitate our understanding of the problem to describe it these various ways.

[22] Cf. Mele & Moser (1997), 233.

[23] Hieronymi, op. cit., 367.

takes to be unwarranted by the evidence she has at hand. Thus, the content of Agnes' self-deceptive intention is not (as Mele would have us suppose) *to believe that Ralph is faithful* but rather something like *to bring it about that she believes that Ralph is faithful*. This may seem a rather unimportant distinction, but it can provide us with a helpful way to describe the kind of activity undertaken by self-deceivers. Agents who are attempting to deceive themselves are not trying to will themselves into believing because belief is not something under their direct voluntary control. But, as we have said, they do have indirect control over their beliefs, and this type of control is something they can intentionally exercise.

Note that belief is not the only outcome at which we can intentionally aim without having direct control over whether or not it actually occurs. Take for example, falling asleep. I can "try" to fall asleep, in the sense that I can form an intention to go to sleep by midnight, and I can engage in activities that will likely result in my falling asleep at this time, but I cannot fall asleep on command. I do not have direct control over whether I fall asleep or not—it is just something that "happens" to me. But, based on my knowledge of what tends to make me fall asleep, I can drink a calming herbal tea or take some Valerian root prior to lying down; I may put on relaxing music or insert a pair of earplugs; I might try to think relaxing thoughts or distract myself from stressful ones; and so on. All of these actions may be intentionally undertaken, in an attempt to increase the chances of success in my falling asleep by midnight. And although my falling asleep itself is not under my direct voluntary control, given what we have just said, it is not in the least problematic to claim that I can be practically committed to bringing it about that I fall asleep by a certain time.

If we don't think it is possible for someone to intend to bring it about that she holds a belief which she believes is unwarranted, this is in large part because we forget that self-deception is almost always a temporally-extended process.. In this respect, it is like many other intentional activities. As Moran and Stone (2009) note, intentional performances are essentially *progressive* in nature—they have *purposive parts* and *unfold* over time.[24] That is, intentional performances have duration, and whether those performances unfold over seconds or minutes, days or weeks, months or years, depends on the nature of the activity in question. Intentionally raising one's hand or typing the word 'action' may require mere seconds, whereas intentionally writing a dissertation or getting one's Ph.D. may take years.

More specifically, intentional activities that we characterize as *projects* are rarely instantaneous or short-lived. Projects are generally complex intentional performances, which require a longer duration of time to carry out—where by "complex", I mean that they are comprised of several "lesser" or "subordinate" intentional activities, which themselves are required as means to the ultimate end in question.[25] Take the example of the agent who intends to get her Ph.D. Achieving this goal is plausibly characterized as an intentional project, the success of which depends upon carrying out several

---

[24] Moran (2009), 143.

[25] I am not trying to provide a strict definition of 'project' here, though I do suspect that having complex intentional subparts is a necessary condition for an activity's counting as a project. Rather, I am appealing to a kind of commonsense spectrum of intentional action with more "basic" intentional actions on one end (e.g., picking up a hammer), slightly more composite actions in the middle (e.g., driving a hammer into a nail), and significantly more complex projects on the other end (e.g., building a birdhouse). Of course, the line between "regular" intentional actions and intentional projects may be somewhat fuzzy, but that is to be expected. Nevertheless, I think we can characterize everything from the lifting of the hammer to the driving of the nail to the building of the birdhouse as intentional, as each of these descriptions represents something done purposively by an agent for reasons that she takes to commit her to the action(s) in question.

subordinate tasks. The agent must attend certain classes and take and pass the relevant exams; she must form a dissertation committee and write her thesis; she must set a defense date and successfully defend the dissertation to the satisfaction of her committee; and so on. Many of this project's sub-parts are themselves are intentional projects, with intentional sub-parts, and so on.[26]

Moreover, the execution of intentional projects need not be strictly temporally continuous to count as enduring. Not everything the agent does during her doctoral studies need be aimed at or part of obtaining her Ph.D., nor need she constantly have her long-term goal in mind to count as still being engaged in that project. She may take an entire semester off or simply go out for beers with her friends; she may temporarily set aside Ph.D. studies in the pursuit of other goals , e.g., raising a family or taking care of a loved one; she may even sometimes completely fail to be aware of the relevant goal, as when she is distracted or unconscious. But none of these discontinuities entail that the agent fails to be engaged in a long-term intentional project to obtain her Ph.D. Indeed, one need not have an explicit intention in mind at every point in one's intentional project, so long as there is some sort of continuous *commitment* (dispositional or otherwise) to the goal in question.[27]

---

[26] I am not going to tackle the problem here of whether every intentional activity can be reduced to some most basic intention or intentional action. This would take us too far afield. However, it does seem right to maintain that intentional activities (especially those we characterize as *projects*) generally have several layers of intentional action which fall under their scope—either as a means to one of the agent's relevant ends (e.g., the ultimate end of the project in question or some end subordinate to that ultimate end) or as a further description under which the action can be said to be intentional. What is important here is that what we call intentional action "is the kind of thing which rationalizes its sub-parts (those actions done 'in order to' do it)" (Ibid., 147).

[27] Of course, given that intentions do represent a kind of commitment to acting, if there are too many discontinuities or significantly long breaks within an agent's so-called project, there

Furthermore, just as the execution of a project need not be temporally continuous to count as intentional, any particular stage of an intentional project need not be temporally preceded by some explicit avowal of the intention relevant to that project or one of its sub-parts. As Moran and Stone note, not every intentional action must be preceded by a "pure intention," i.e., when the agent "intends to do something but hasn't yet done anything else in order to do that."[28] As we have seen, intentions represent *commitments* to acting, and such commitments are often expressed in the actions which represent their implementation, not in a pure mental state of the explicit form, "[Next, shortly, tomorrow, Monday…] I am going to φ"—though they may sometimes take this form.[29] Many of the intentional actions we repeatedly undertake—especially those subordinate intentional actions regularly employed toward achieving some further end—over time, require less that one form a pure intention prior to or simultaneous with one's undertaking the means; rather, they tend to occur more as a matter of habit. Moran and Stone give the examples of rolling out of bed

---

will likely come a point at which we are inclined to say the agent is not really committed to the goal in question, or that she has abandoned the project, or that she fails to have the goal in some other way—i.e., that she does not *really* intend to bring about what she proposes.

[28] Ibid., 142.

[29] Intentions will likely take this form in cases where making them explicit is useful or important to the adopting of the intention in question or to the execution of the proposed action. For example, a severely depressed agent may need to form the explicit intention, "I am going to get out of bed today," in order to get herself to undertake the action of getting out of bed. Similarly, someone who has just learned how to make a complicated cocktail may need to (either mentally or verbally) list the steps needed in order to successfully make the drink in question, such that they may form pure intentions of the sort, "First, I am going to pour 2cl of vodka into the glass; then 3 oz. of sweet vermouth," and so on. Or, in cases where the agent needs some sort of reassurance, she may form a pure intention like, "I am going to win this race" prior to stepping up to the starting line. Additionally, pure intentions are often expressed in order to communicate our intentions to others: "Today I am going to clean my house"; "later I am going to the movies". In all of these cases (and others, I am sure), the intention is pure, in the sense that the agent has not yet undertaken any means toward bringing about the intended end, but (as we have said) the pure intention must still represent a *commitment* to doing some particular action—not a mere belief (or, e.g., a mere hope) regarding what one is going to do—to count as an intention at all.

in the morning and changing speed according to traffic.[30] Similarly, to take an earlier example, it may be that when one is learning to drive a car with a standard transmission, one forms a pure intention before upshifting or downshifting because doing so is difficult and requires effort. However, as one becomes accustomed to changing gears, one may not expressly form the pure intention, e.g., "to shift into third gear" before one does so. That is, performing the action of shifting into third no longer "exist[s] apart from the things one does."[31] But this is not to say that one does not shift intentionally. Rather, it is only to claim that one need not affirm an explicit intention prior to the initiation of an action.[32]

Thus, although "[a]ny action of significant duration is apt to have moments of pure intending and pure acting among its innumerable parts,"[33] it need not have either at any particular point in time to continue to count as intentional under some description.[34] An agent may possess an intention (in the sense of being practically committed) to bringing it about that he believes a certain proposition, and he may pursue this aim, even if he does not consistently dwell on that intention or make his reasons explicit to himself.

---

[30] Ibid., 144.

[31] Ibid.

[32] I think the same may be said for an agent's awareness of her *reasons* for acting as she does. Agents may be in possession of and act on reasons, without constantly reflecting on them or explicitly reaffirming their commitment to them (though, as with intentions, they may do so).

[33] Ibid.

[34] Furthermore, it is important to note that, just as the execution of intentional projects need not be preceded by pure intentions, neither must they terminate in so-called "perfected intentions"—that is, in action successful in achieving its goal. As we have said above, whether an intention is "perfected" in action may depend both on contingent factors outside the agent's control and on the agent herself and her continuing commitment to pursuing the aim in question.

To claim that one need not consistently dwell on one's intentions or reasons for engaging in a particular action is not to insist that some intentions are inaccessible to the agent who has them. On the contrary, I think it an important requirement on intentional action that an agent's commitments to and reasons for action be at least *available* to her for reflection. If an agent "acts" on a particular intention and for reasons to which she herself has no access, it is difficult to see how such "action" is voluntary and under her control. Indeed, it is difficult to understand how being moved by such intentions qualifies as *acting*. For it is difficult to understand how the person with such an intention can make sense of what she does as *her* answer to Anscombe's "why" question. Of course, this is not to say that our reasons need always be fully transparent to us at every moment, nor that we always need to know which of our (often competing) reasons actually ends up producing the action in question. Rather, the claim here is that for an agent to act intentionally, she must have some degree of access to her reasons for so acting.

However, this returns us to the second problem regarding intentional self-deception we raised above—namely, that an agent's awareness of her intention to try to bring it about that she believes a certain proposition threatens to destroy her self-deceptive project altogether. Indeed, even if we think it possible for an agent to attempt to bring it about that she believes a certain proposition, it seems that a requisite condition for her succeeding is that she be *un*aware of her intention to do so. Otherwise, the project appears self-defeating. How are we to deal with this worry?

First, it is worth considering that agents often undertake intentional actions and projects which they do not think they can successfully execute. An agent may adopt and pursue an end without thinking that she will actually attain it. Because of the contingency

of the outcome of many of our projects—i.e., the dependency of success on a large number of external factors—we may still attempt to bring about certain consequences, even if we think we are likely to fail. I can attempt to flip an omelet in the pan with a single flick of my wrist, knowing full well that I am not an accomplished chef and that similar attempts in the past have failed miserably. Likewise, I may try to hit a ball over the Green Monster at Fenway Park or make a half-court shot without having any significant amount of confidence that I actually can do so. I may even try to become President of the United States, knowing full well that my chances of winning a national election are very slim.[35]

However, the problem facing us in the case of intentional self-deception is not simply that the agent thinks it unlikely that she will succeed in deceiving herself. That is, it is not merely a practical problem of an agent's lacking confidence in having the requisite ability to succeed in her endeavor, but also a conceptual problem raised by the nature of belief itself. If it is constitutive of belief that it aims at the truth,[36] and yet one is aware that one is trying to acquire a belief one takes to be false (or at least takes to be unwarranted by the evidence), this makes self-deception appear self-defeating from the get-go. I may attempt to flip an omelet or make a half-court shot, knowing full well that I will likely fail, because I may recognize that I have good reasons for making the attempt (e.g., becoming a better cook or trying to win a thousand dollars). And there is always the

---

[35] It is generally accepted that one cannot intend something one thinks it impossible to do. Yet one can *try* to do these things. We might, therefore, recharacterize the above intendings as "intending to *try*," or, as Michael Thompson suggests, "intending combined with confidence in success"—in this case, a low level of confidence. Cf. {{85 Thompson, Michael 2008/s103;}}.

[36] Cf., for example, Velleman (2000).

slim chance that I might succeed—even if the odds are against me. But if I *recognize* that I am trying to get myself to believe something I take myself to have no good epistemic reason to believe, the nature of belief itself seems to preclude any chance of success. As Ariela Lazar (1999) puts it, "the [prudential] goal of holding the desired belief often conflicts with the [epistemic] goal of making it true."[37]

It is important to note, however, that self-deceptive projects do not need to take deceiving oneself or acquiring a false belief as their ultimate goal. In fact, acquiring or maintaining a false belief for its own sake seems rather absurd.[38] Rather, it seems fairly obvious that the acquisition or maintenance of he belief in question will be conducive to some other end(s) or goal(s) the agent has, where these provide the agent with the prudential reasons she takes herself to have to deceive herself. Thus, self-deceptive projects will generally (as with may of the intentional projects we undertake) represent subprojects in the service of another of the agent's aims. But what might such an aim be? And does a commitment to this end require that one be committed to self-deception in a way that would allow us to characterize it as intentional?

4.6 Self-Deception and Self-Image.

We mentioned briefly above that one of the reasons Agnes might have for deceiving herself is to maintain a particular image she has of herself. We all adopt various practical identities in our day-to-day lives (for example, I may identify myself with the role of daughter, philosopher, feminist, American, baseball fan, and so on), and

---

[37] Lazar (1999), 273.

[38] We can, for sure, imagine an agent who tries to acquire a false belief "just to see if he can," but such a case would be quite unusual.

some of these practical identities are those with which we very strongly identify—those that, in some sense, make up our "core" self.[39] When these identities come under fire—when the stability of our self-image is threatened—we tend to feel a certain level of anxiety or cognitive unease. This is clearest when we notice that our own actions and inclinations suggest that we are not the way we think we are, and not the way we wish to be. For example, though I value myself as a self-sufficient individual, I may find myself forming unhealthy dependencies on other people. Or I may believe I am a good friend yet find that my actions suggest that I tend to treat my friends thoughtlessly, or even unkindly. Similarly, an agent may strongly wish to be a certain kind of person yet find himself with certain inclinations that do not bear this projected self-image out—as when an extremely conservative Christian believes strongly that homosexuality is a sin, and thus places extreme worth on not being gay, yet finds himself consistently desiring members of the same sex. In other cases, some fact about the external world may threaten one's view of oneself. For example, in the case of Agnes above, we may say that her image of herself as a beloved wife, as someone deserving of not being lied to, etc., whose husband would never lie to her, is threatened by her assessment of the evidence that Ralph is being unfaithful.

As Kant writes in *Religion Within the Limits of Reason Alone*, "man is never more easily deceived than in what promotes his good opinion of himself."[40] And, as we have seen, such threats to self-image may, in many cases, motivate a kind of wishful thinking

---

[39] I do not mean to put forward an ontological thesis regarding the nature of the self here. Rather, I am concerned more with the ways in which agents view, regard, or otherwise value their identities in the sense that is important for agency and action.

[40] Quoted in Witschen (2008), 139. My translation.

or other biasing mechanism, which may in turn produce a false or otherwise unwarranted belief in the agent. This likely explains why many agents overestimate their abilities, believing that they are smarter, stronger, or otherwise "better" than they really are. The analysis offered by Barnes and Johnston may be apply to such cases, insofar as the self-deception involved may reduce anxiety regarding the self. It may even serve a broader evolutionary function associated with self-preservation—assuming that such overconfidence is (or was historically) conducive in some way to species survival.

It is important to note that one's self-image is not a purely passive, static set of beliefs about oneself. Although in one sense one often simply "finds" oneself with particular beliefs about oneself (beliefs that may arise due to introspection, observation of one's behavior, or from one of the aforementioned biases), one's self-image is also something dynamic—something that must be cultivated and maintained. One may even speak of one's self-image as part-reflection, part-projection—what one observes or believes oneself to be like, and what one wishes (desires, hopes, wants) oneself to be. When these two sides conflict, or when one has a tendency to overidentify oneself with one side or the other, one has a strong motive to engage in self-deception.[41]

It seems quite obvious that one can undertake intentional projects of self-improvement. I may desire to be physically fit and embark on a diet; I may wish to be drug-free and check into rehab; I may value being educated and enroll myself in a continuing education program. In all these cases, I desire certain of my practical identities (e.g., being overweight, being "clean", being ignorant) to be other than they are, and I

---

[41] I think something like this is what Sartre has in mind in his discussions of *mauvaise foi*—in which an agent may be said to identify too strongly with either her facticity or her transcendence. Cf., for example, Sartre (1956), 48ff.

may undertake an intentional project aimed at changing these facts about myself. Similarly, I may undertake projects of self-improvement with the goal of changing my character. I may notice that I treat loved ones unfairly or that I have a tendency to be self-centered, and I may engage in certain activities (e.g., trying to be nicer, thinking more about other people, and so on) in an attempt to acquire the virtues of character I value. All of these types of projects involve a concerted effort on my part to develop in myself a certain practical identity that I would like to embody. But one may also intentionally engage in projects aimed at *maintaining* one's current practical identities. I may value my identity as a soccer fan or as a fluent German speaker, but without watching any of the World Cup or regularly speaking German, I may find it difficult to identify myself with such a person. Thus, I may undertake certain intentional projects (e.g., watching World Cup soccer or regularly attending a German-speaking group in town), in order to further cultivate or maintain my self-image—and I can do so without being subject to the charge of irrationality.

Just as one can intentionally engage in perfectly rational projects aimed at self-improvement and –maintenance, it should not seem so strange that one can also embark on a project of self-deception in the service of one of these goals. However, as we have seen, self-deceptive projects are not as easy to understand as the type of endeavors we have just discussed—for self-deception requires that, rather than *becoming* or *remaining* a certain kind of person, one come to *believe* (or persist in believing) that one is a certain way, and this appears to require that one not be aware of one's intention to acquire or maintain this belief . Yet we have also said that for an action to count as intentional, one must have reflective access to one's intentions. How are we to deal with this problem?

It is, perhaps, not necessary to insist that a self-deceiver must be completely *unaware* of her deceptive intention.. An agent need not completely repress her intention to acquire or maintain a certain belief in order to intentionally attempt to achieve this goal. Of course, if her attempt is to be successful, she must usually *distract* herself from reflecting too heavily on her deceptive intention—e.g., she must selectively direct her attention, think positive thoughts, and so on—but, as we have seen, this is something that agents can and actually do regularly accomplish in the service of perfectly rational projects. Many of our projects require that we distract ourselves from thinking too hard about our intentions in order to succeed. In order to fall asleep at a certain time an agent may need to think thoughts not having to do with sleeping or trying to sleep. Other tasks, like hitting a home run or running 5 kilometers in under 20 minutes, too, may also be hindered by thinking too much about what one is attempting to do. So we should not be all that surprised to find that self-deceivers engage in similar types of psychological manipulation in the service of their self-deceptions.

4.7 The Irrationality of Self-Deception: Acting to Acquire Reasons

Self-deception differs from the other projects we have discussed in being irrational. It is important to see that my account does justice to this fact, even as, on this account, the self-deceiver undertakes an intentional project aimed at coming to believe a certain proposition for prudential reasons having to do, say, with the preservation of her self-image. Self-deception is a form of irrationality because it involves intentionally doing something one has good reason to believe cannot be done: bringing it about that one has sufficient epistemic reason to believe a given proposition, without doing anything to change the facts. In cases of rational belief acquisition and/or maintenance, an agent

already has evidence at hand that he takes to justify a particular belief. In contrast, the self-deceptive agent must generate epistemic reasons for believing something she currently thinks it would be irrational to believe. This involves a deeper kind of irrationality characteristic of the motivationally-based beliefs we discussed in Chapter Three—the agent's irrationality consists of the fact that she attempts to violate her own rational standards.. As Julius Schälike writes:

> In paradigm cases, we therefore appear to speak of self-deception when the person attempts, *with deliberate intention*, to generate a belief that she, at *this* point in time, *takes to be false*. When the person also employs strategies, which, from her perspective, do not operate in the service of a reality-oriented belief-forming process but rather of "epistemic sabotage," then we are presented with self-deception – even if the person characterizes this process *ex post [facto]* as a "learning process."[42]

Schälike aptly notes that should the self-deceived agent actually acquire the belief in question, she will likely no longer view the acquisition of this belief as irrational. She may characterize her self-deceptive project as having been a "learning process" like any other. But, as he points out—and as we have seen in our discussion of intentional self-deception thus far—agents engaging in self-deception are not engaged in empirical hypothesis testing, or some other "reality-oriented" method of acquiring epistemic reasons. They are engaged in a process of "epistemic sabotage," intended to produce reasons that will make it psychologically possible for them to believe what they currently take to be unwarranted. Belief is, by its very nature, constitutively aimed at truth. Thus, self-deceiving agents transgress the very norms necessary for them to be said to be epistemic agents in the first place.

---

[42] Schälike (2004), 374. My translation.

Of course, the individual engaging in self-deception does not thereby cease to act in her capacity *qua* agent. She is still reasons-responsive, in both a pragmatic and epistemic sense. She employs means to her end, insofar as she looks for reasons to believe, and she does so for prudential reasons. But she realizes that she cannot bring herself to believe for non-epistemic reasons, so she must generate said reasons if she is to achieve her goal. Thus, in some sense, we may call such an agent "pseudo-rational," for she does not cease to care about reasons. Yet she engages in a type of activity that violates the norms of reasons-responsiveness—for good epistemic reasons are not the type of reasons that can simply be *generated* via the strategies she employs.

This sort of project occupies what Pears calls "the territory of cognitive dissonance," insofar as an agent who attempts to acquire or maintain a belief in ways that violate her own epistemic standards experiences a certain level of psychic tension or discomfort.[43] Indeed, the measures we mentioned above (e.g., selective evidence gathering, directing one's attention, rationalization, acting "as if," and so on) are all means of attempting to resolve or avoid the kind of cognitive dissonance involved in attempting to believe something one takes oneself to have little to no good epistemic reason to believe.

I hope it has become clear that the account of self-deception I have provided in this chapter is not only plausible (insofar as it explains the phenomena we commonly observe among those individuals we take to be self-deceived), but also that it is superior to the other accounts we have examined thus far. I now want to put to one side questions about the nature of self-deception, and consider instead, whether self-deceivers always warrant criticism. Is it always wrong to deceive oneself? If not, what distinguishes the

---

43 Pears (1982), 279.

exceptional cases from the others? It is these and other related questions regarding the morality of self-deception that I wish to discuss in the final chapter of this work.

CHAPTER 5. "YOU OUGHT TO KNOW BETTER": SELF-DECEPTION AND MORALITY

5.1 Self-Deception and Epistemic Responsibility

"It is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence."[1] With these famous words, W.K. Clifford introduced an evidentialist principle regarding the norms of belief—one William James would later criticize in his essay, "The Will to Believe".[2] On the face of it, Clifford's principle does seem a bit strong, especially given that human beings appear to be cognitively "hardwired" to believe beyond (or even in opposition to) the weight of the evidence, as we have seen in previous chapters. The prevalence of cognitive biases like those we discussed in Chapters Two and Three often affect our belief systems without our being aware that they are doing so. And if this is so, it seems ignorance may excuse agents who believe against the evidence without knowing it. Such agents are, through no fault of their own, not in the right epistemic position to believe "responsibly," such that they do not appear to be directly accountable for the possession of at least some of their unwarranted beliefs.

However, we have seen that even non-intentionalists like Mele and Barnes claim that "self-deceived" agents who acquire or maintain certain beliefs via a motivationally biased treatment of the evidence may be responsible for their irrational beliefs, assuming they were, in some sense, "competent" to avoid so believing. Thus, even biased agents may be epistemically culpable for believing as they do, insofar as they are guilty of a kind of "epistemic negligence"—a failure to attend to the evidence in a non-biased fashion, despite possessing the requisite ability to do so. Of course, this attributes a rather weak sense of responsibility to self-deceived agents. Self-deceivers are merely

---

[1] Clifford (1879), 186.

[2] James (1979).

"careless" or "slipshod" regarding the evidence for their beliefs. Such agents "ought" to know better—but they simply fail to exercise certain capacities in their possession, with the result that they do not know better. Thus, on such views, to say that one is culpable for one's unwarranted beliefs is simply to say that one has failed to keep a certain bias in check, or to carefully examine the reasons one believes as one does, or to "watch out" for biases that lead to unwarranted beliefs. Now surely epistemic agents are (at least sometimes) responsible for their beliefs in this way—at least if we assume that they are to some degree *aware* of their tendencies to treat evidence in a motivationally biased fashion. That is, if one knows that one is subject to certain biases, and one subsequently fails to keep a watchful eye out for the operation of these biases on one's beliefs, then it does seem that one may be epistemically blameworthy for failing to do so. In fact, the mere fact that one has certain types of biases in the first place may represent a deficiency in one's epistemic character that needs to be remedied or repaired. Yet, as we have said, this is still a rather weak sense of epistemic responsibility, for we are only accusing biased believers of failing to know better, not of *actually* knowing better.[3]

On the other hand, if we adopt an intentionalist, diachronic account, such as the intentional-project account of self-deception we explored in the last chapter, we may be able to postulate a stronger sense of accountability for self-deceivers. If self-deceivers are capable of intentionally undertaking projects of self-deception, they may be more than merely negligent—they may be responsible for actively sabotaging their evidence. In other words, if we are right about self-deception, then self-deceivers are not merely

---

[3] This account of epistemic responsibility also raises certain puzzles: How are we to understand the "competence" to avoid one's biases? Do Mele, et al. merely mean that, counterfactually, had an agent paid attention to the operation of a certain bias within her, she would not have been deceived? If so, we run the risk of making the agent's responsibility for her self-deception appear trivial. Furthermore, if agents really are able to resist said biasing mechanisms, why do they fail to do so? Are they weak-willed? Do they willfully avoid exercising these capacities? If so, a level of intentionality might creep into the picture on this level that would make the non-intentionalist somewhat uncomfortable.

inattentive or careless; not only *ought* they to know better—they *do* know better. Furthermore, self-deceivers may not only be guilty of intentionally manipulating the nature of the evidence regarding the particular proposition in question; they may also be guilty of sabotaging their evidential standards in general, which may lead to the acquisition of what we might call "epistemically vicious" character traits that may affect their future belief-forming processes. Thus, if agents can straightforwardly act to violate their own epistemic standards, perhaps self-deception represents a kind of epistemic failure for which agents are blameworthy in the fullest sense of the word.

However, Clifford's principle not only introduces an epistemological norm regarding belief, it also implies that believing on insufficient evidence is *morally* wrong. (Hence the essay's title: *The Ethics of Belief*.) And we may pose a related question regarding self-deception. It seems quite obvious that self-deception generally represents some kind of epistemic failure. But is it also an moral failure? Is self-deception always, everywhere, and for anyone *morally* wrong? For, in some sense, self-deceivers are both the perpetrators and the victims of their deceptions. As Herbert Fingarette writes: "[The] 'epistemological' paradox [of self-deception] generates moral paradox since ignorance and blindness exculpate, whereas knowledge, insight and foresight inculpate."[4] Of course, if (as we said in the last chapter) one's ignorance or epistemic blindness regarding one's practical identity or a particular state of affairs related to that identity is the result of one's practical commitment (i.e., intention) to bring about that blindness, then it seems an agent may be no more a "victim" of self-deception than a suicidal agent who throws himself in front of a moving train is a "victim" of manslaughter. At the same time, there is the possibility that self-deception may lead to objectively better consequences for the agent or those around her. Some theorists assert that certain "vital lies" or "positive

---

[4] Fingarette (1969), 136.

illusions" regarding oneself may be important or even necessary for one to successfully act in and/or interact with others in the world.[5] Perhaps self-deception may lead to more happiness for all concerned, or to an improvement in one's moral character, or to a preservation of one's sense of autonomy. If this is so, then it might also be possible that self-deception is, in some cases, morally permissible or even praiseworthy.

In the rest of this chapter, I wish to explore the question of the morality of self-deception in greater detail. I will attempt to show that engaging in self-deception does, in fact, generally represent a kind of moral failure on the part of the agent. I will argue that, even though in some cases individual acts of self-deception may lead to better consequences for all concerned or to a kind of reaffirmation of autonomy, self-deception in general threatens these desired outcomes, such that adopting a policy of self-deception is either irrational or just plain wrong, regardless of whether one adopts a character-based, deontological, or consequentialist approach.

5.2 Self-Deception and Moral Character

We said in the last chapter that intentional projects of self-deception are often (if not always) undertaken for prudential reasons having to do with one's self-image. When an agent's valued self-image is threatened by the recognition that the evidence points to a certain fact's obtaining, be that fact about her character (e.g., that she is a bad mother, or a procrastinator, or a coward, or a racist) or about states of affairs in the world that directly relate to her self-image (e.g., that her love for another is unrequited, or that her colleagues do not like her, or even that others are being persecuted around her while she does nothing), she may resort to intentional strategies (e.g., rationalization, selective attending and evidence-gathering, etc.) aimed at maintaining that self-image. We have also seen that the more entrenched one becomes in one's self-deception, the more

---

[5] See, for example: Taylor (1989).

habitual such behavior becomes and the less difficult it becomes to "fool" oneself. One becomes used to rationalizing away and generating alternative explanations for negative evidence, with the result that one is significantly more resistant to the force of potential counterevidence.

In a moral sense, self-deception aimed at the maintenance of a particular false self-image is doubly dangerous. First, it may lead to precisely that at which it aims—to an unwarranted belief about one's self—making it possible in many cases for one to mistake vices for virtues, undesirable character traits for desirable ones. Second (and perhaps more dangerously), it may lead to the acquisition of what we might call "epistemic vices"—character traits relating to one's epistemic agency that threaten to undermine one's ability to interact with oneself and the world in a reality-responsive manner. Habitual self-deception undermines an agent's commitment to pursuing truth, and may thus lead to a kind of unwillingness (or even inability) to be objective in one's assessment of certain states of affairs in which one has a vested interest. And this degradation of the agent's epistemic character threatens to alienate her not only from herself (in the sense that it leads to a lack of self-honesty and self-knowledge) but also from the world around her (including threatening her relationships with other agents).

Thus, the acquisition of certain epistemic vices may result in the acquisition of morally vicious traits (and vice versa).[6] For example, as Barbara Ehrenreich points out, insistence on self-deceptive positive thinking may produce a kind of habitual overoptimism or underpreparedness. She gives the example of many government officials prior to 9-11, whose inordinate emphasis on optimism and positive thinking led them to

---

[6] I will remain neutral on the question regarding whether epistemic and moral vices are to be *identified* with one another, or whether the one type of vice is causally *derivative* of the other. I am inclined, however, to say that in some cases, the epistemic vice may itself represent a moral vice (as with the undermining of autonomy, which I discuss below), whereas in other cases one type of vice may be said to emerge as a result of the activity of the other type (e.g., moral cowardice may lead to epistemic cowardice, or vice versa).

grossly overlook rather obvious warning signs of an impending terrorist attack on American soil. She writes: "That fact that no one…heeded these disturbing cues was later attributed to a 'failure of imagination.' But actually there was plenty of imagination at work…there was simply no ability or inclination to imagine the worst."[7] We can quite easily envision that such an inability or disinclination to draw rational conclusions from the evidence might be the result of prolonged engagement in self-deception, where this kind of self-deceptive activity results in the degradation of one's moral character. Another stark example is that of regular German citizens during World War II who willfully "looked the other way" when the SS deported millions of Jews, Gypsies, homosexuals, communists, and other minorities to concentration camps located just outside their communities. Some of these citizens claimed to not have known that gross moral atrocities were being committed more or less in their own backyards, though their behavior indicates that many of them were engaged in a kind of active self-deception regarding the immoral treatment of prisoners by the Nazis. This, in turn, led to an mutually-supportive unwillingness to act, which, as history has shown us, had dire moral consequences.[8]

However, one might object that agents may also engage in self-deceptive projects as a partial means of developing certain moral *virtues*. An agent may value being a friendly, social person but realize that she more often acts cruelly and antisocially—and as a means to acquiring the valued character trait, she may try to deceive herself into believing that she is friendly and social, in the hope that this particular piece of self-deception will lead to her actually becoming this way. Might this use of self-deception (in

---

[7] Ehrenreich (2009), 10-1.

[8] This example also aptly demonstrates the importance of other-deception and reciprocal support or affirmation to certain projects of self-deception. In many cases, the willingness of others to "play along," "yes-man" each other, mutually affirm each other's claims, and so on may be necessary to support a particular individual's or group's self-deceptive projects.

the service of virtue acquisition) be morally commendable? I think it is not. While deceiving oneself may, in some cases, help one more easily engage in behavior that contributes to one's acquiring a certain virtue or morally commendable character trait, the acquisition of this virtue will not be clear-eyed. Part of making moral progress in building one's character is acquiring knowledge not only of the virtue to be cultivated, but also of whether or not one has that virtue. And although self-deceptively believing that one embodies a certain virtue may even eventually lead to one's having that virtue, it will do so only accidentally. Furthermore, there are other, more commendable ways of acquiring virtuous character traits that do not involve self-deception, as I will discuss shortly.

But first it is important to notice that engaging in self-deception, even in an attempt to become a "better" person, still threatens one's reality-oriented perspective toward the world by making one less responsive to truth-conducive evidence or by altering one's evidential standards in ways that make one less likely to have true beliefs regarding one's character. And this may threaten one's integrity as a moral agent in general. If one becomes skilled in deceiving oneself in order to acquire a virtue, that same "ability" may be just as easily employed in the service of vice.

Second, an agent who is entrenched in a self-deceptive project aimed at bettering her character will likely not be able to view her actions with clear eyes. Although she may perform the right actions (i.e., those a virtuous agent would perform), as Russ Shafer-Landau (channeling Aristotle) notes, acting continently is to be distinguished from acting virtuously. Virtues are more than just mere patterns of behavior: "People are virtuous only when their understanding and their emotions are well integrated. A virtuous person who understands the right thing to do will also be strongly motivated to do it, without regret or reluctance, for all the right reasons."[9] Yet the self-deceiver is by no

---

[9] Shafer-Landau (2010), 247.

means "well integrated." Indeed, as we have seen, the very project of self-deception ensures that she is not. Thus, it is not clear that the self-deceived agent can be motivated by the right sorts of reasons, given that her reasons get their motivational force from a self-deceptive belief that she is virtuous, when in fact she is not. So long as she engages in self-deception, she lacks a kind of understanding of herself that prevents her from being truly virtuous. She remains in a very important sense alienated from herself in a way that does not allow for the acquisition of the type of fully-integrated "second nature" required for one to be said to possess the virtue in question—and this threatens to make the end of her self-deceptive project (i.e., the acquisition of the relevant virtue) even more difficult to attain.

Yet one might wonder whether this emphasis on self-honesty and truth-centered rationality leads to an overintellectualized and perhaps even destructive obsession with having true beliefs and avoiding cultivating false ones—to a life "devoid of vigor and the satisfaction of many important needs and desires…[that] render[s] us emotionally impoverished by constraining us always to submit to hard facts."[10] Nevertheless, as I mentioned briefly above, there appear to be other ways of developing virtuous character traits that are morally preferable to relying on self-deceptive techniques. While engaging in self-deception allows the agent to (at least temporarily) put out of her mind the fact that she is not as she wishes to be (with the supposed aim of eventually becoming what she wishes to be), there are, as Mike Martin points out, more reasonable ways of "transcending available evidence in forming beliefs, attitudes, and emotions."[11] He lists three such ways, namely faith, hope, and imaginative expression. I wish to discuss each of these phenomena briefly here, in an attempt to demonstrate why these ways of going

---

[10] Martin (1986), 126.

[11] Ibid., 127.

beyond the evidence are morally preferable to self-deception, at least as regards the development and betterment of one's moral character.

Martin defines *faith* as "a belief not based on evidence establishing it as true or even on a belief contrary to the main direction of evidence available to a person." He goes on to maintain that faith involves "active" belief that demonstrates a kind of *trust* in what is believed.[12] Of course, faith may sometimes involve a level self-deception (and thus represent a kind of "bad faith"), but it might not: "For although it entails going beyond the evidence in forming beliefs, it does not necessarily involve evasion of evidence, truth, or self-acknowledgment of how things appear to be."[13] In his discussion of William James, Martin claims that many of our most significant beliefs outstrip the evidence we have for them, and among these are many of our moral and religious beliefs and so-called "self-confirming or self-verifying faiths—beliefs that if acted upon make it more likely that they will become true."[14] One may trust in the fact that one will become a more friendly person (or, in cases where the evidence is inconclusive regarding one's current state, that one *is* a friendly person), and this may directly contribute to one's becoming a friendlier person, without engaging in the types of self-deceptive projects we discussed in Chapter Four. And the adoption of such beliefs might be more morally commendable toward the end of acquiring a particular virtue than their self-deceptive counterparts, for in "good faith", one still keeps a watchful eye on one's beliefs. One does not avoid, rationalize, or otherwise evade evidence, but one maintains a healthy confidence in those forced, momentous, live options, while at the same time taking care not to slip into self-deception. This is no mean feat, and good faith may quite easily revert

---

[12] Ibid.

[13] Ibid.

[14] Ibid., 128.

to a kind of self-deceptive bad faith, but this should not be surprising. There is no reason to suppose that self-improvement or virtue acquisition is easy. Thus, the diligent agent with her heart set on bettering her character may embrace a kind of faith in her moral improvement without engaging in self-deception, but she must always be careful that she is keeping a watchful eye on her faith and its relation to the evidence.

However, there are other ways of self-honestly outstripping the evidence that involve regarding a proposition as true without thereby actually straightforwardly believing it. Take, for example, cases of *hope*. Martin describes hoping that *p* as involving believing *p* to be possible (yet not certain) and living as though *p* will occur.[15] It is more than merely wishing, insofar as one can be said to wish for things both when one believes they are impossible or highly unlikely and when one lives as though those things will not occur.[16] And it differs from faith, insofar as faith involves an level of trust or certainty in the relevant proposition, whereas hope does not. In one sense, being hopeful may require acting "as if" (e.g., acting as if one were friendly), and we have seen that such pretense before oneself and others may contribute to one's self-deception. Indeed, hope, like faith, may be both the result or the source of certain self-deceptive projects. However, just as pretense is not always deceptive, hope, too, may be expressed non-self-deceptively. The rational hoper is clear-eyed about a) the realistic possibility of *p*, b) the fact that *p* is not yet the case, and c) the fact that her living as though *p* will occur does not mean it will. Thus, the hopeful agent trying to cultivate a the virtue of friendliness will still remain realistic about her factual limitations, her current status as, e.g., not-always-friendly, and the fact that her acting as though she is or will be friendly

---

[15] Ibid., 129.

[16] For example, I can wish that I was immortal (which I take to be impossible) or that I will become a billionaire (which I take to be highly unlikely), and I can still be said to wish these things even if I do not live as though I believe they will someday obtain.

might not make it so. Thus, although she must remain careful that she does not slip into self-deception, the hopeful agent regards the desired state of affairs to be true in a different way than that of full-blown belief or faith. And such clear-eyed hope seems more likely to produce overall improvements in one's moral character than the workings of self-deceptive mechanisms.

Finally, *imaginative expression* includes ways of apprehending oneself and the world such as daydreaming, fantasizing, willingly suspending disbelief, pretending, joking, experiencing something vicariously, and so on.[17] In all of these cases, one exhibits and/or experience emotions toward objects one takes to be unrealistic, nonexistent, or in some other sense "imaginary." I can daydream that I have completed writing this chapter or fantasize that I am being interviewed on the Colbert Report; I can suspend my disbelief that there are no such things as demons and yet allow myself to be scared by "The Exorcist," or I may pretend to be possessed myself; I can joke to another that I am a "typical blonde" when I do something stupid; I can even "enjoy" watching you eat ice cream as though I were eating it, even though I am on a diet. All of these situations involve regarding or presenting propositions as true within a particular imaginative context we take to be unreal(-istic). Yet, as David Velleman notes, none of them require that the agent regard these propositions as true "in a manner designed to reflect whether it really is true."[18] That is, such imaginative expressions differ from straightforward beliefs, insofar as they are cognitive attitudes "in which a proposition is regarded as true without concern for whether it really is."[19] Like beliefs, imaginative

---

[17] Cf. Ibid., 130. Note that imagination may, in some cases, be conducive to one's maintaining a sense of hope, as described above.

[18] Velleman (2000), 112.

[19] Ibid.

expressions regard a proposition as true, but unlike beliefs, the agent does not take such cognitive attitudes to actually reflect the reality of what their content embodies.

Similar to faith and hope, imaginative expressions may provide fertile soil for the cultivation of self-deception, but they need not. One constructive way for an agent to self-honestly develop a virtuous character trait like friendliness may be to imagine herself acting in a friendly manner—to pretend, to fantasize, to imagine being on the receiving end of friendliness, and so on. Such imaginative exercises can be immensely helpful in the training of one's character—so long as one does not succumb to the wiles of one's imagination, mistaking fantasy for reality or willfully allowing one's imaginings to control one's assessments of reality. Thus, just as with faith and hope, the imaginative agent must remain watchful, but such a way of regarding the world still seems superior to self-deception as a means toward developing positively one's character.

The irrationality constitutive of self-deception brings with it cognitive tension and unease, undue resistance and repressiveness, a sense of being at odds with oneself—but exercising faith, hope, or imagination is by no means irrational and need bring no such discomfort. The self-honest agent regards her character in a clear, open-eyed fashion, but this does not preclude her from entertaining desirable propositions and regarding them as true in a way that may aid in her bettering her character; neither does it require that the agent slip into some kind of internal irrationality to achieve this result. Thus, it appears that even in cases where self-deception is employed as a means to acquire a particular (moral) virtue of character, there are other, more rational ways to achieve such virtue without engaging in the type of irrationality constitutive of self-deceptive projects. Self-deception is, therefore, not to be adopted as a rational policy toward the betterment of one's character. Not only does it often result in a degradation of one's reality-responsiveness toward oneself and the world, it may limit one's clear-eyed access to one's moral reasons for acting in ways that the aforementioned cognitive attitudes do not.

Moreover, as I have intimated above, self-deception may threaten to undermine one's status as a moral agent altogether. This, of course, does not bode well for one's moral character, but it also points to another way in which self-deception may represent a moral failure: it may threaten the stability of an agent's autonomy. And if we agree with theorists like Kant that autonomy is the ground of all moral action, it would seem that self-deception might correspond to more than a mere potential cause or effect of moral vice; rather, it might turn out to be *the* central moral vice. Or, to move away from the terminology of virtue theories and toward that of deontological theories, self-deception might be said to violate one of the central duties a person may be said to have—a duty which grounds the possibility of moral action in the first place: namely, a duty to oneself. It is to a discussion of self-deception in this context that I now turn.

5.3 Self-Deception, Autonomy, and Duties to Oneself

Two interrelated elements central to Kant's ethical theory are, of course, the notions of autonomy and respect for persons. For Kant, autonomy is construed not only negatively as the capacity to act free from constraints external to our will (which also include such "internal" motivations as emotions, desires, impulses, drives, and so on) but also positively as the capacity for free, rational choice, i.e., the capacity for self-legislation and -determination (which includes the ability to set our own ends and to direct our pursuit of these ends). Autonomy is what endows us with the "Humanity" that distinguishes *persons* from other types of beings, including non-rational beings that act in goal-directed ways (as with, say, children or non-human animals). Furthermore, it is this capacity that make persons inherently worthy of respect, as we see embodied in Kant's Formula of Humanity, which directs us to treat persons (in the form of ourselves and others) never merely as a means to some further end but also always as ends in themselves.

Insofar as I recognize myself to be an autonomous individual, then, I have a duty to myself to treat my Humanity as an end in itself (that is, to respect myself and my capacity for autonomy, for moral self-determination), and the highest form such respect can take is to aim not merely at self-preservation but also at moral self-perfection. Nelson Potter writes:

> The idea that there are duties to oneself is basic to the structure of Kant's ethical thought, including his well-known emphasis on the importance of moral freedom. [Duties to oneself have] a common characteristic: a significant relation to the goal of preserving, maintaining, developing, and perfecting the very centre of our being as human beings, our moral self.[20]

Since, on Kant's ethical theory, one's moral autonomy is what makes one worthy of respect in the first place, to fail to respect one's Humanity in the form of one's moral freedom is to threaten one's ability to act in one's capacity as a moral agent. Failure to adhere to duties toward oneself, then, threaten to undermine one's moral agency altogether. However, the preservation, maintenance, and development of our moral self entails that one have the ability to engage in moral self-constraint. Part of what it is to be a *finite*, *imperfect* being capable of autonomous moral action is to be subject to temptation—to being determined from without, to heteronomy. Hence, all moral duties require a degree of self-restraint and are thus all "partially [duties to ourselves] because the agent must use the powers of self-constraint that are presupposed by any duty to recognize and undertake any duty at all."[21]

Furthermore, Kant's motivational internalism, combined with his claim that the truly moral action is performed not only in accordance with duty but also *from duty*, leads

---

[20] Poetter (2002), 377-8.

[21] Ibid., 376. Note that this also implies that morally perfect rational beings would have no duties to themselves, given that they have no need of moral constraint, due to the perfect alignment of their incentives and their will. In other words, since morally perfect beings can never be tempted, they know no duties to themselves. (Cf. Ibid., 378-9.)

to the identification of justifying and motivating reasons in cases of morally right action. Yet in cases of heteronomous action, these reasons come apart. An action that should be performed for duty's sake is motivated instead by a contingent desire, emotion, or other inclination. To put it a bit differently, moral action is necessarily autonomous action, given that it is grounded in our moral freedom.

For the above reasons, we can understand why, in the *Metaphysics of Morals*, Kant writes that the first command of all duties to oneself is "*know* (scrutinize, fathom) *yourself*."[22] He specifies the type of self-knowledge with which moral self-perfection is primarily concerned in one of his lectures on ethics:

> The individual has a universal duty regarding himself to so dispose himself that he should be capable of observing all his moral duties and thus be able to establish in himself moral purity and moral principles and strive to act according to them. This is thus the first duty to oneself. *Now, to this duty belong introspection and self-exploration as to whether or not one's dispositions also have moral purity*.[23]

For Kant, knowing whether one is motivated by duty or by some inclination external to the will is central to the duty of moral self-perfection. We cannot make moral progress if we are habitually blind to the source of our motivations, or if we cannot recognize tempting inclinations as such, or if we continually mistake heteronomous for autonomous action. We can thus see why self-deception represents a serious violation of the fundamental duty to oneself, insofar as it hinders one's ability to introspect, to self-examine, to scrutinize and fathom oneself—in other words, to *know* oneself.

Potter points out that there are two broad types of self-deception that directly threaten one's moral self-knowledge (and thereby one's moral agency) on Kant's view. The first type concerns the strength of one's moral versus one's personal motives.

---

[22] Kant (996), 191.

[23] In Witschen (2008), 137. My translation and emphasis.

Assuming as Kant does that 'ought' implies 'can' and that "it is a transcendentally valid presupposition that the strength of the moral motive is always adequate to the morally required action," agents often deceive themselves into believing that their personal motives are stronger than they really are, or into denying that the strength of the moral motive is adequate to the action in question.[24] In cases such as these, agents may deny that they are strong enough to do as morality requires, or view themselves as more or less "determined" by their contingent desires—as when a married man exclaims to his lover, "I have lost all control! I cannot but continue to sleep with you, even though I know it is wrong." While it may sometimes be true that an agent simply does not have the capacity to act autonomously in a particular situation (as with cases of physical or psychological compulsion), the cases of interest here are instances in which an agent *can* but does not *want* to submit to the demands of morality (which he himself has recognized and thus placed on himself as a duty). Thus, instead of respecting his Humanity and exercising his autonomy, he attempts to convince himself to believe that he lacks the strength of will required to employ his moral freedom. Instead, he rationalizes that his personal motives outweigh his moral motives in strength and thus that, in some sense, he "cannot" do as morality requires.[25]

The second type of self-deception centrally related to Kant's proposed duty to know oneself, Potter claims, is when an agent deceives herself that her motive is a moral motivation (i.e., from duty) when her motive is really from self-love or some other inclination. In such cases, the agent does not try to convince herself that she cannot do as

---

[24] Potter, op. cit., 387.

[25] There is an interesting parallel here to Sartre's notion of "bad faith" in cases where one denies one's transcendence (i.e., radical freedom) by overidentifying with one's facticity (over which one has little to no control). In the above case, the adulterer identifies himself with his desire for his lover (which represents a brute fact about him), causing him to self-deceptively deny his freedom (to do as he perceives morality to require) and enter into bad faith. Similar ideas can be found in Kierkegaard as well.

morality demands. Rather, she attempts to convince herself that she is, in fact, acting autonomously when she is really acting from some other self-interested motivation or inclination. Thus, we can imagine the unfaithful husband above ending his affair and returning to his wife, and he may do so under the pretense that "it's the right thing to do," when in fact he returns solely out of guilt or fear or cowardice. Here, the self-deceived agent fails to know himself by denying his true motivations. Instead of self-deceptively denying his ability to act autonomously, he self-deceptively overasserts it. He willfully fails to recognize his true motivations and thus still violates Kant's fundamental duty of self-knowledge.

Of course, in both of these types of cases, the "self-deception" in question might merely amount to something like wishful thinking or motivated biasing, in which an agent is simply mistaken about the strength or nature of her motivations. In such cases, attributions of responsibility to agents who violate Kant's dictum to "know thyself" might be more akin to those of the kind of epistemic "negligence" we discussed above than to something for which they are fully blameworthy. In other words, we *ought* to try to discover our true motivations, but, as Kant himself notes:

> The depths of the human heart are unfathomable. Who knows himself well enough to say, when he feels the incentive to fulfill his duty, whether it proceeds entirely from the representation of the law or whether there are not many other sensible impulses contributing to it that look to one's advantage (or to avoiding what is detrimental) and that, in other circumstances, could just as well serve vice?[26]

So perhaps even Kant wants to say that the impossibility of our ever knowing our true motivations is part of what causes us to err regarding the true strength or nature of our motivations. Of course, if Kant really does adhere to the claim that 'ought' implies 'can', it would seem that self-knowledge must be, at least to some degree, possible. However,

---

[26] In ibid., 387.

Potter takes Kant to be saying here that our imperfect self-knowledge—our inability to delve into the "unfathomable depths" of our hearts—is the *result* of self-deception, not its origin.[27] Indeed, we might think that it is not because we *cannot* see into our hearts that we fail to know ourselves, but rather because we *do* not.

In *Religion within the Boundaries of Mere Reason*, Kant maintains that all agents have developed an inclination toward evil and thus incorporated evil into their maxims. In this sense, we are all "fallen", and our self-knowledge is thereby limited. However, as Potter notes, for Kant, "the major, perhaps the sole source of evil is *self-deception*, the inner lie, by which we defeat morality in us, and thereby defeat ourselves."[28] And this seems to be self-deception in an active, not a merely passive sense. Thus, in a way similar to the virtue theorists above, human beings may be responsible for their epistemically ignorant states (and the resulting moral failures that derive from this ignorance), insofar as they at some time actively participated in the self-deception that led to a degradation (and, in Kant's case, crippling) of their moral agency. At any rate, it is clear that for Kant, self-deception represents a moral failure to the highest degree, given that self-knowledge is part and parcel of the most fundamental duty human persons have—a duty to themselves. Of course, many philosophers (even those of a deontological bent) have claimed that Kant's view is too extreme, especially as regards his injunctions against lying (be it a lie to another or an "inner lie" to oneself).[29] But what if we adopt a softer approach—one centered around duties to respect persons in general? Is it possible that

---

[27] "Our failure is a failure of self-knowledge. And that failure is because of our defeating ourselves through self-deception" (Ibid., 388).

[28] Ibid., 386.

[29] There are, of course, further worries about Kant's approach, including the possibility of committing willful evil action in the first place, but to discuss these issues would take us too far afield. My goal thus far has merely been to show that self-deception represents a gross moral failure on Kant's own proposed account.

self-deception might sometimes enforce and affirm one's agency instead of undermining it?

Instances of guilt amongst concentration camp survivors represent a very interesting case of such potentially "autonomy-constructing" self-deception. It is often difficult to explain why many survivors of the Holocaust experience guilt for, through no direct fault of their own, having lived where others died. One possible explanation for this behavior is that, while interred in the concentration camp, almost every facet of their autonomy was threatened. Prisoners were treated as sub-human, and were continually subjected to public and private humiliation. They were often arbitrarily punished or even executed for made-up infractions (or sometimes merely for fun, or to strike fear into other prisoners). And such prisoners likely needed to psychologically cling to whatever facets of their autonomy they could hold onto to survive. Thus, when reflecting after the fact on their having survived, they might self-deceptively attribute more agency to themselves than they actually possessed at the time and actively cultivate a sense of guilt over having survived. This may also explain why survivors cling to such guilt, for to remove it would be to take away the last semblance of agency or personhood that person views themselves as having had at the time.[30]

However, such self-deceptive guilt, while preserving a sense of *past* agency in memory, may have dire consequences for the agent in her present circumstances insofar as it may severely threaten her current ability to function as an autonomous agent. Her ability to adequately relate to her past, to her current self, and to others may be undermined by her self-deceptive guilt, such that self-deception in these cases, while maintaining an agent's present sense of past autonomy, still may represent a kind of disrespect for her current person. Indeed, self-deceptively taking *too* much responsibility

---

[30] I am grateful to Gry Ardal Christiansen for the wonderful discussion that led to this example.

for one's actions can also represent a kind of disrespect for oneself, insofar as one fails to respect one's limitations or (as Sartre would put it) one's "facticity."

But what about cases in which self-deception is employed to preserve or regain a current or future sense of self-respect? What about the case of Agnes in Chapter Four, who attempts to retain an image of herself as loved and respected wife by engaging in self-deception regarding her husband's faithfulness? We have already said that many if not most cases of self-deception are aimed at a kind of preservation of self-image, and we can imagine that Agnes' self-deceptive project represents an attempt to maintain some sense of self-respect. Yet there are still two very important senses in which her project itself may be said to represent a moral failure regarding respect for persons. First, if she tries to bring it about that she believes that Ralph loves and respects her, despite her recognition on some level that he does not, she is disrespecting herself by allowing a belief in a falsehood to determine her sense of self-worth. She does not treat herself as a person worthy of not being deceived, either by Ralph or by herself. Second, she fails to respect Ralph's moral agency by failing to hold him responsible for his infidelity. Part of respecting persons involves viewing them responsible for their actions, yet by engaging in self-deception regarding his actions, Agnes lets her husband "off the hook," as it were, and this is also to disrespect his status as an autonomous agent.

There are, of course, borderline cases, e.g., of severely depressed agents who need to tell themselves that their life has meaning or that they are loved to get out of bed in the morning, where doing so serves as a kind of crutch—a way to maintain any sense of autonomy. (The same may be true of slaves, prisoners, and other agents whose autonomy is severely limited by external psychological or physical forces.) These are tricky cases, insofar as the agency of such individuals is already severely limited or impaired. Of course, here it is also important to note that self-affirmations directed at boosting one's sense of autonomy need not be self-deceptive. Just like fantasies and imaginings, positive thinking, pep talks, and the like need not involve an active attempt to believe the

proposition(s) in question, yet they may be employed to help bolster an agent's sense of self-respect or autonomy. I may tell myself repeatedly that I will finish an article before the deadline, in order to boost my confidence, while still remaining open to the possibility that I will not finish on time. Likewise, a depressed agent may give herself a pep talk to help her get through the day, without attempting to actually get herself to believe it. The mere thought (even if combined with thoughts to the contrary) may be enough to allow her to assert her autonomy and maintain a minimal sense of self-respect without her engaging in any sort of agency-undermining irrationality. Of course, the line between positive thinking and self-deception may be blurry here, but so long as the agent is not actively engaging in agency-undermining self-deception, her activities may still be morally permissible or at least excusable.

A final point is in order here concerning the assessment of moral responsibility of self-deceivers from a duty-based perspective centered around respect for persons. Insofar as such theories are focused on the motivation of the agent and the nature of the action in question, the consequences of an agent's self-deception must remain largely at the wayside. Whether her particular project of self-deception ends up actually leading to an affirmation of her autonomy or to a greater respect for persons in general is largely irrelevant on such a view. What is wrong with self-deception is not that it tends to lead to bad consequences, but rather that a lack of respect for persons is embedded in the very nature of the self-deceptive activity itself. And this has to do with the character and motivation of the self-deceiver, not with the eventual results of her deception. Aside from threatening her very ability to act as an agent in the world, self-deception reflects a lack of concern on the part of the self-deceiver, either for herself or for others, and thus even cases of self-deception that (contingently) lead to an affirmation of autonomy are rooted in a project that itself is morally suspect.

However, one might think that consequences are of greater importance to the moral evaluation of self-deception than that put forward by the above theories. Of course,

in one sense, we have already shown that self-deception tends to lead to "bad" consequences—even when viewed from the perspective of those theories. On traditional character-based moral theories that laud traits like courage, temperance, lovingness, justice, humility, generousness, and so on, self-deception leads to a lack of integration within the agent that threatens her ability to acquire the virtue in question or to be said to act virtuously at all. On duty- and autonomy-based theories, self-deception endangers one's fundamental moral agency and often leads to a failure to properly respect persons. Yet what about normative theories according to which the goodness or badness of an action is determined by the amount to which the action results in pleasure, the satisfaction of desires, or some other notion of subjective or objective well-being?[31] Might not self-deception lead to good consequences in this sense? Many social psychologists argue that a level of self-deception is beneficial or even necessary to maintaining a healthy mental life and/or self-image. I wish to conclude my discussion of the morality of self-deception by addressing this claim. I will argue that, even based on these types of consequentialist considerations, an agent would do better to adopt a policy of self-honesty than one of self-deception.

5.4 Self-Deception and Happiness.

In her influential book, *Positive Illusions: Creative Self-Deception and the Healthy Mind*, Shelly E. Taylor argues that self-deception in the form of "positive illusions" can contribute to one's overall well-being and mental health.[32] However, as both Mike Martin and Neil Van Leeuwen note, Taylor tends to conflate self-deceptive

---

[31] Such theories include (but are perhaps not limited to) both egoistic and utilitarian accounts. On egoistic theories, what will matter is the nature of the consequences for the agent herself, whereas on utilitarian theories, the emphasis will be on the overall well-being of all agents concerned.

[32] Taylor (1989).

illusion and mere hopeful optimism or imaginative expression.[33] In some places, she goes so far as to equate illusion with mere false belief, but as Martin points out, many of our false beliefs are the product of unconscious mechanisms and cognitive biases, and we have argued that these arational processes do not amount to self-deception. He writes: "to the extent that unconscious processes amount to routine brain processing, they constitute self-deception only using a very broad sense [of the term] that includes all false beliefs that our brains play a role in generating."[34] Thus, Taylor's claim that self-deception is beneficial rests on a rather shaky understanding of the concept of self-deception.

Nevertheless, Taylor's observations linking self-deception and happiness are not entirely without merit. Her claim is more descriptive than normative, and even if we agree with Martin and Van Leeuwen that having positive illusions is not always the same as being self-deceived, we can plausibly imagine that such illusions may at least sometimes be the result of a self-deceptive project. So the normative question remains: If having positive illusions about ourselves contributes positively to human happiness, might it not be the case that it is sometimes morally permissible (if not required) to pursue policies aimed at fostering such illusions? And if one such policy involves cultivating and engaging in self-deceptive projects, then might not a policy of self-deception, too, be morally commendable?

Of course, the notion of 'happiness' is a difficult one, and numerous theories of value have been put forward regarding just what this term is supposed to encompass. Nevertheless, in this final section, I will attempt to show that self-deception is *not* generally conducive to happiness, where happiness is understood as a kind of flourishing

---

[33] Cf. Martin (1986), 40; Van Leeuwen (2009), 116. I will lean heavily on these articles in what follows.

[34] Martin, op. cit..

or overall human well-being, and thus it should not be adopted as a policy of action.[35] By "overall human well-being," I mean the combination of both a kind of subjective well-being (e.g., a function of desire-satisfaction, pleasure or something of the like) and a kind of objective well-being (e.g., having certain worthwhile external goods, living a life judged to be valuable by certain external standards and values, and so on). As Martin writes: "Subjective and objective well-being are [usually] interwoven, and they are bridged by the sense of meaning that is so important to happy lives. … [A]lthough our sense of meaning is subjective, it puts pressure in the direction of justified values, and truthfulness about them, for most of us attempt to ground our sense of meaning in such values."[36] Van Leeuwen puts forward a similar view. He claims that although we might be able to achieve a kind of "Matrix happiness," consisting of mere positive feelings but lacking in genuine external goods, this does not constitute "choiceworthy happiness," which consists of both subjectively pleasurable sentiments and the possession of objectively worthwhile goods.[37] Thus, in what follows, I am envisioning a rather broad theory of value that encompasses both these subjective and objective components of human happiness.[38]

Before I begin, however, a few words are in order on the distinction between an *instance* and a *policy* of self-deception. On virtually any consequentialist theory, an individual act of self-deception may, in some cases, lead to better actual consequences

---

[35] There are, of course, theories of value that claim that 'happiness' is a function of the extent to which one, e.g., fulfills one's potential as a rational being. Such teleological theories will not be the focus of this section, as I think it quite obvious, given what we have said above, that self-deception is not conducive to happiness in this or any similar sense.

[36] Ibid., 40.

[37] Cf. Van Leeuwen, op. cit., 109-10.

[38] There are many versions such a theory of value could take, but I do not have the time to take them all up here. However, I am reasonably certain that my arguments will apply generally to theories of the type described above.

than the alternatives. But in this chapter, we have concerned ourselves with the evaluation of self-deceived *agents*: Does self-deception represent a moral failure on the part of the agent herself? Can agents be morally justified in pursuing a project of self-deception? Of course, on theories according to which what makes an action right or wrong is a matter of the goodness or badness of its actual consequences, some instances of self-deception will likely be "right," all things considered—but this judgment can only be made after the fact (if it can ever be made at all). We are interested in the point of view of the deliberator—for even if a particular instance of self-deception may turn out *ex post facto* to have been the right thing to do, we might still think the decision to pursue a self-deceptive project represents a moral failure on the part of the agent, especially if doing so has a tendency to lead to worse consequences than not.[39]

So what might a policy of self-deception aimed at happiness look like? Van Leeuwen lists four main elements such a policy would embody:

1. Awareness of areas in one's life that are felt to be lacking or about which one has insecurities.[40]
2. On the basis of awareness of the sort mentioned in (1), [explicit or implicit] selection of which beliefs will promote happiness by working against the tendency toward [the] negative affect the awareness engenders.
3. Commitment to attending to information that seems to confirm the beliefs selected in (2).
4. Commitment to ignoring evidence that disconfirms beliefs selected in (2).[41]

---

[39] Van Leeuwen makes a similar point in ibid., 111.

[40] We might rephrase this condition as an awareness of a dissatisfaction with or insecurity regarding one's self-image, as discussed in the previous chapter.

[41] Ibid., 113.

Note that this type of self-deceptive policy appears compatible with the intentional-project account of self-deception we developed in Chapter Four. Yet can such a policy lead to an agent's happiness?

Taylor argues that overinflated opinions of oneself raise one's self-esteem. That is, believing you are a better driver or student or parent than you really are makes you subjectively happier. However, it is not clear whether this correlation between overestimation and happiness, if real, shows that self-deception actually leads to happiness, even if we assume that a policy of self-deception may lead to such "positive illusions." As Van Leeuwen points out, it is perhaps more plausible that having high self-esteem (or being happier) causes the false beliefs one has about oneself and not the other way around. He writes:

> If beliefs that attribute positive features cause positive affect, we'd expect not to see so many intelligent or good-looking or successful people who are unhappy. But there are many such people. But if positive affect causes self-flattering beliefs, then it's possible for one to have beliefs that attribute positive features to oneself without having positive affect, where these beliefs arrived by another route. Furthermore, we'd still expect to see a correlation between positive affect and self-flattering belief, which is precisely what we find.[42]

He also notes that when agents make overoptimistic judgments about themselves, they may simply be *interpreting* their "positive moods."[43] Thus, it is not clear that self-deception aimed at producing Taylor-esque positive illusions leads to happiness, even if it turns out that positive illusions and happiness are in some way correlated. However, even if we accept the above argument, we have not yet shown that self-deception does *not* tend to lead to more happiness than its alternatives. However, I think an argument can be made for this thesis.

---

[42] Ibid., 118.

[43] Ibid., 119.

First, if we agree with Van Leeuwen that there is a distinction to be made between "Matrix" and "choiceworthy" happiness (hereafter, MH and CH, respectively), it seems that self-deception leads primarily to the former and not to the latter type of happiness. Indeed, it appears to create stumbling blocks in our attempts to satisfy our desires: "Having true beliefs allows people to accomplish things… [whereas] false belief tends toward dissatisfaction."[44] But self-deception tends toward false beliefs, which makes self-deceivers likely candidates for having their desires thwarted. The champion of self-deception may object here that, since the self-deceiver in some sense "wants" to believe that the world is a certain way, deceiving herself may actually lead to the satisfaction of her desires. However, here it is important to note that the driving motivation behind most agents' self-deceptions is not a desire to *believe* that the world is a certain way, but rather a desire that the world *actually be* that way. And having false beliefs about the way the world really is does not satisfy this central desire. The self-deceived wife does not actually make it the case that her husband is not cheating on her merely by believing it. She does not actually *have* a faithful husband; she only believes she does. Thus, self-deception appears to only lead to MH, not CH.

Of course, it may be true that some cases of MH may lead one to come to possess certain worthwhile external goods that might transform one's MH into CH. Pleasurable or positive sentiments may breed the type of attitude required for the acquisition of certain objective goods, as we see in the case of friendship. Unfriendly, overly-negative people are unlikely to be able to make genuine friends, whereas positive, sociable feelings may help one in establishing such relationships. Thus, although one may self-deceptively cultivate positive sentiments about oneself and/or others without actually having any friends (MH), the having of such sentiments may assist one in making friends,

---

[44] Ibid., 120.

thereby leading to a more choiceworthy happiness. And if self-deception can help one achieve these kinds of positive sentiments, why not think that this type of self-deceptive MH might also be conducive toward cultivating CH?

Here, it is first important to note that it will be very difficult for self-deceivers to tell when MH is conducive to CH and when not. The former may in some cases contribute to the latter, but the connection is by no means a necessary one—and given that false beliefs tend to undermine desire-satisfaction, agents cannot rely on self-deception to successfully and consistently breed positive sentiments within themselves. Furthermore, as we have noted in the above sections, engaging in self-deception tends to inhibit one's truth-responsiveness in general. This makes it even more difficult for self-deceivers to know when policies of self-deception may be beneficial and when not. Van Leeuwen gives the illustrative example of a father who is self-deceived about his son's intelligence:

> On complicated matters, like the intelligence of a child, there will be ways in which one can contribute and be helpful, as well as aspects of the situation one cannot change. To know the difference between the aspects of the child's mind one can help and the aspects one cannot, one has to be responsive to the evidence the child provides of his abilities. A policy of self-deception is deliberately contrary to such responsiveness. One is likely to end up being self-deceived not just about the native abilities of the child, but also about abilities one could help improve. In short, self-deception, even on the assumption it ever *could* be helpful for happiness, undermines awareness of the conditions for its own helpfulness.[45]

Thus, even though self-deception may sometimes lead to a kind of MH that may be conducive to CH, it still should not be adopted as a policy on the part of the agent desiring CH, for self-deception tends to lead away from the satisfaction of one's motivating desires and may even undermine one's ability to pursue CH in the first place.

---

[45] Ibid., 124.

However, perhaps one can deceive oneself regarding whether or not one is "truly" happy (i.e., whether or not one has CH). Such a type of self-deception would result in a kind of MH, in which the agent is deceived regarding her "true" happiness, in the sense of CH. And one might argue that, in this sense, MH is really all that matters for human happiness. While it may be difficult to deceive ourselves regarding whether we are subjectively experiencing pleasure or pain (though this, too, may be possible), given our normative sense of happiness as represented by CH, it seems very likely that one can become self-deceived regarding whether or not one is happy in this latter sense. Martin discusses the case of Tolstoy's Anna Karenina, who leaves her somewhat boring husband for an exciting new lover. At first, this brings her joy and exhilaration, but after a while, the relationship proves to be shallow and ultimately unfulfilling. Furthermore, she misses her son (whom she abandoned to be with her lover). Thus, Anna finds herself in an unhappy situation, yet she attempts to deceive both herself and others into believing that she is happy. Why does she do so? Martin writes: "One reason is that acknowledging her unhappiness risks intensifying her misery…. A second reason is that she senses that acknowledging her unhappiness would confront her with the necessity of making a decision about how to proceed with her life—or how not to proceed, as she moves toward suicide."[46] So perhaps in such a case, a life of MH, in which Anna Karenina deceives herself that she has CH, is preferable to a life in which she is forced to admit that she has made serious mistakes, that she has pursued temporary pleasure at great personal cost (to herself and others), and that, ultimately, she is unhappy.

Yet this does not seem quite right. As we have seen, cognitive tension and dissonance tends to be characteristic of self-deception and self-deceptive projects. Such a state is not desirable, yet it is difficult to avoid. Even if she is successful in her self-

---

[46] Martin, op. cit., 34.

deceptive endeavors, whatever Matrix-type happiness Anna Karenina might achieve via self-deception is a temporary one. MH as a result of self-deception is unlikely to be stable or enduring, given that the agent likely often must confront evidence to the contrary. As Van Leeuwen writes, self-deceptive MH "tends in the direction of its own undoing."[47] He quotes a study by Colvin, Block, and Funder, which claims that a "deep albeit perhaps unrecognized and unacknowledged sense of uneasiness consequently may pervade the self-[deceiver], hardly a condition conducive to mental health."[48] For example, Anna will have to wonder why her belief that she is truly happy does not quite match up with her emotions or actions (e.g., experiencing feelings of disappointment, taking nightly doses of morphine, and so on). Moreover, such an uneasiness often "seeps through" one's demeanor, making it likely that the self-deceiver will be treated differently by others. Anna's friends and relatives may suspect she is really unhappy and thus act differently toward her than they might toward a happy person, e.g., by pushing her to return to her husband and son, or by refusing to spend time with her, or something of the sort.

Thus, the instability of self-deceptive MH may (and likely will) lead to worse consequences than adopting a policy of self-honesty. Even if her misery is intensified by the fact that she must make a decision regarding how to proceed with her life from this point, the admission that she is unhappy opens up possibilities for Anna that self-deception closes off to her. From a utilitarian standpoint, Anna is now in a position to take responsibility for her actions to others—to repair her relationship with her son, to make things right, and thereby to potentially increase the happiness of others. From a purely egoistic standpoint, recognizing her unhappiness allows Anna to integrate and understand herself and her actions—to recognize herself for what she is. This may be a

---

[47] Van Leeuwen, op. cit., 121.

[48] Quoted in ibid., 122, n.20.

difficult and unpleasant realization at the time, but in the long term it is likely to lead to more happiness than remaining in the tense and unstable state of self-deceptive MH. Just as addicts often realize that taking responsibility for the hurtful and unkind actions they committed while under the influence is an important step on the road to recovery, self-deceivers like Anna Karenina may find that she can only become the person she desires to be by first being honest with herself and others. Thus, even in this sense, self-deception does not seem likely to lead to enduring and stable happiness.

5.5 Conclusion

For all these reasons and others I do not have time to discuss here, I hope to have plausibly demonstrated in this chapter not only that agents may be both epistemically and morally responsible for their self-deceptions, but also that an agent's adopting a policy of self-deception is not rational and represents a kind of moral failure, regardless of the normative ethical approach one adopts. Self-deception makes the agent susceptible to epistemic and moral vices, threatens to undermine her sense of integrity and autonomy, compromises her ability to interact in a reality-responsive way with herself and others around her, and tends to lead to more overall unhappiness than happiness. While positive thoughts and attitudes may be beneficial to an agent's mental health, there are better ways to foster such sentiments in oneself (in both a practical and a moral sense) than by engaging in self-deception, as with self-honest instances of faith, hope, and imaginative expression. As we have seen, although self-deception itself is an unstable and sometimes self-defeating type of intentional activity, it nevertheless often appears easier to take on such an irrational project than to be honest with ourselves. Yet the easiest alternative is rarely the moral one, and so it is in this case. Self-deception in the service of immorality has been a crucial component in many (if not all) of the greatest atrocities committed throughout history, and we must be careful to not allow ourselves and our society to become willing victims of similar deceptions. Being a responsible believer is no mean

feat, yet I conclude this chapter with an exhortation to the reader to exercise diligence, to keep a watchful eye on her beliefs and belief-forming processes, and to whenever possible avoid engaging in self-deception. Self-honesty and self-knowledge are difficult to attain, but it is something we ought to strive for—for the betterment of both ourselves and those around us.

EPILOGUE

This dissertation has focused on questions regarding the metaphysical and psychological possibility of self-deception and has attempted to show that self-deception is a phenomenon best characterized as both motivated and intentional, with the result that self-deceivers can be held responsible for their deceptions in a stronger sense than that of being merely epistemically negligent.

In Chapter One, I introduced the static and dynamic paradoxes of self-deception, which arise when one attempts to draw a close analogy between self- and other-deception. I used the example of Parker the Parkinson's denier to show the various ways in which his unwarranted belief might be characterized, from mere ignorance or psychological compulsion to wishful thinking to weak, non-intentional self-deception to strong, intentional self-deception. Since the latter kind of account presupposes there is a direct analogy to self-deception, I went on to show how the various ways one understands interpersonal deception may mirror the various accounts one might give of strong self-deception. I concluded the chapter with a brief discussion of the role of empirical studies in philosophical investigations of irrationality.

Chapter Two introduced a series of intentionalist accounts of self-deception, namely theories that divide or otherwise partition the mind to make sense of the so-called "internal irrationality" of the self-deceiver. I discussed both the stronger, Freudian versions of this theory and the weaker versions put forward by Lockie, Pears, and Davidson. Employing the example of Agnes, the self-deceived wife, I concluded that both types of divided-mind accounts appear to raise more metaphysical worries than they solve and thus do not adequately make sense of the phenomena. I also attempted to show that if there is such a thing as mental compartmentalization or Freudian-type divisions within the mind, an account of self-deception centered around such divisions will not be of the intentionalist persuasion.

In Chapter Three, I moved on to a discussion of non-intentionalist accounts of self-deception, specifically those of Alfred R. Mele and Annette Barnes. On such theories, self-deception is said to be motivated but not intentional, such that agents need not have intended to deceive themselves for them to count as self-deceived. Such theories have gained in popularity in recent years, due to their claim to be explanatorily more parsimonious and in-line with empirical investigations of supposedly irrational belief-forming processes. Against these theories, I argued that there are certain phenomenon we take to be central to self-deception that Mele, Barnes, et al. cannot account for, e.g., the cognitive tension observed in self-deceived agents, the reflective nature of self-deception, the pseudo-rationality of self-deceivers, and so forth. I thus proposed that a more robust account of self-deception is necessary to make sense of these phenomena.

Chapter Four puts forward just such an account. I claimed that if we focus more heavily on the diachronic process by which self-deceivers elicit and/or maintain their beliefs over time, what emerges looks much more like an intentional project aimed at the manipulation of one's evidence or evidential standards than a mere more-or-less unconscious process of motivated biasing. I suggested that if we view self-deception in the way I propose, we can escape the paradoxes of self-deception, while at the same time making sense of the features lacking on non-intentionalist accounts. I discussed intentions and intentional action in more detail, in order to explain just how it is that self-deception can be said to be intentional, claiming that agents engaged in self-deception have as their end a practical commitment to develop or maintain a particular self-image. But rather than taking means to actually become or remain a certain way, self-deceivers instead attempt to bring it about that they *believe* they are that way. This involves engaging in a type of deliberate, pseudo-rational epistemic sabotage, and it is this that makes the self-deceiver epistemically irrational.

Finally, in Chapter Five I argued that self-deceivers are not only epistemically but also morally responsible for their self-deceptions, in a sense stronger than that allowed

for by a charge of mere epistemic negligence. Self-deceivers are actively and willfully involved in undermining their evidential standards and thus their status as epistemic agents, and this threatens their status as moral agents as well. I showed that self-deception exhibits a tendency to cultivate both epistemic and moral vice in an agent, to undermine her autonomy and respect for persons, and to lead to worse consequences than other, more responsible ways of entertaining propositions such as those involved in faith, hope, and imaginative expression. For this reason, I concluded my discussion of self-deception with the claim that self-deception generally represents a moral failure on the part of the moral agent.

This work represents an attempt to show not only that intentional self-deception for which an agent bears responsibility is metaphysically possible, but also that it is plausible to think that something like this provides a better explanation for self-deceptive behavior than the other theories discussed here. While each of the theories I reject likely point to very real phenomena, my claim is that it is philosophically beneficial to make the conceptual distinctions I have drawn here and to claim that such phenomena represent something other than self-deception. For example, non-intentionalist theories point to a very common type of weak irrationality, namely our tendency to be cognitively or motivationally biased in favor of certain propositions. And partitioned-mind theories point to the fact that the mind is a complex system that is not always perfectly integrated. Nevertheless, neither of these theories makes conceptual space for the possibility of a stronger kind of internal irrationality, despite the fact that we are the types of creatures that can, in fact, undertake intentional projects in the service of such irrationality. It is this that I have attempted to demonstrate here. Making these distinctions also points us to some interesting avenues for future research. For example, what is the relationship between interpersonal deception and self-deception? Does the success of our self-deceptive projects tend to depend on the complicity of others, their willingness to play along, and so forth? Is it possible for groups of people to be collectively self-deceived?

What role do emotions like shame and guilt play in motivating and sustaining self-deception? Might the presence of such emotions indicate an empirical distinction between mere biasing and self-deception proper? These are all interesting questions, and I hope this work will encourage future research into these areas.

REFERENCES

Abelson, Raziel. 1977. *Persons: A Study in Philosophical Psychology*. New York: St. Martin's Press.

Anscombe, G. E. M. 1957. *Intention*. 1st ed. Oxford: Basil Blackwell.

Bach, Kent. 1981. An Analysis of Self-Deception. *Philosophy and Phenomenological Research* 41 (3) (Mar.): 351-70.

Barnes, Annette. 1997. *Seeing Through Self-Deception*. Cambridge: Cambridge University Press.

Bermúdez, José Luis. 2000. Self-Deception, Intentions, and Contradictory Beliefs. *Analysis* 60 (4) (October): 309-19.

———. 1997. Defending Intentionalist Accounts of Self-Deception. *Behavioral and Brain Sciences* 20 (1): 107.

Buss, Sarah. 1997. Weakness of Will. *Pacific Philosophical Quarterly* 78 : 13-44.

Cavell, Marcia. 1998. Beside One's Self: Thinking and the Divided Mind. *Crítica: Revista Hispanoamericana De Filosofía* 30 (89) (Aug.): 3-27.

Clifford, William Kingdon. 1879. The Ethics of Belief. In *Lectures and Essays of W.K. Clifford*, eds. Leslie Stephen, Frederick Pollock. Vol. 2. London: Macmillan.

Davidson, Donald. 2004. *Problems of Rationality.* Oxford: Oxford University Press.

———. 1998. Who is Fooled? In *Self-Deception and Paradoxes of Rationality*, ed. Jean Pierre Dupuy, 1-18. Stanford, CA: CSLI Publications.

———. 1985. Deception and Division. In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, eds. Ernest LePore, Brian P. McLaughlin, 138-148. Oxford: Blackwell.

———. 1982. Paradoxes of Irrationality. In *Philosophical Essays on Freud*, eds. Wollheim, Richard and James Hopkins, 289-305. Cambridge: Cambridge University Press.

Dupuy, Jean Pierre. 1998. *Self-Deception and Paradoxes of Rationality*. Stanford, CA : CSLI Publications.

Ehrenreich, Barbara. 2009. *Bright-Sided: How the Relentless Promotion of Positive Thinking Has Undermined America.* New York: Metropolitan Books.

Elga, Adam. 2005. On Overrating Oneself... and Knowing It. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 123 (1/2, Selected Papers from the 2003 Bellingham Conference) (Mar.): 115-24.

Fingarette, Herbert. 1969. *Self-Deception*. Atlantic Highlands, NJ: Humanities Press.

First, Michael B., Allen Frances, and Harold Allen Pincus. 2002. *DSM-IV-TR handbook of differential diagnosis*. Washington, DC: American Psychiatric Publishing, Inc.

Gardner, Sebastian. 1993. *Irrationality and the Philosophy of Psychoanalysis*. New York: Cambridge University Press.

Grunbaum, Adolf. 2001. Does Freudian Theory Resolve "the Paradoxes of Irrationality"? *Philosophy and Phenomenological Research* 62 (1) (Jan.): 129-43.

Gur, Ruben, and Harold Sackeim. 1979. Self-Deception: A Concept in Search of a Phenomenon. *Journal of Personality and Social Psychology* 37 (2): 147-69.

Hamlyn, D. W., and H. O. Mounce. 1971. Self-Deception. *Proceedings of the Aristotelian Society, Supplementary Volumes* 45: 45-72.

Hieronymi, Pamela. 2009. The Will as Reason. *Philosophical Perspectives*: 201-20.

———. 2008. Responsibility for Believing. *Synthese* 161: 351-73.

———. Reasons for Acting (Draft), http://www.humnet.ucla.edu/humnet/phil/faculty/Phieronymi/PHonAction.htm.

Hirstein, William. 2005. *Brain fiction : Self-Deception and the Riddle of Confabulation*. Philosophical Psychopathology. Cambridge, MA.: MIT Press.

———. 2000. *Self-Deception and Confabulation*. Vol. 67. East Lansing, MI: Philosophy of Science Association.

James, William. 1979. The Will to Believe. In *The Will to Believe and Other Essays in Popular Philosophy.*, eds. Frederick H. Burkhardt, Fredson Bowers and Ignas K. Skrupskelis, 1-31. Cambridge, MA; London: Harvard University Press.

Johnston, Mark. 1988. Self-deception and the nature of mind. In *Perspectives on Self-Deception*, eds. Brian P. McLaughlin, Amélie Oksenberg Rorty, 63-91. Berkeley, CA: University of California Press.

Kant, Immanuel. 1996. *The Metaphysics of Morals*. Trans. Mary Gregor. Cambridge: Cambridge University Press.

Krizan, Zlatan, and Paul D. Windschitl. 2009. Wishful Thinking About the Future: Does Desire Impact Optimism? *Social and Personality Psychology Compass* 3: 227,227-243.

Kunda, Ziva. 1990. The Case For Motivated Reasoning. *Psychological Bulletin* 108 (3): 480-98.

Lazar, Ariela. 1999. Deceiving Oneself or Self-Deceived? On the Formation of Beliefs "Under the Influence". *Mind* 108 (430): 265-90.

———. 1997. Self-Deception and the Desire to Believe. *Behavioral and Brain Sciences* 20 (01): 119.

Levy, Neil. 2008. Self-Deception Without Thought Experiments. In *Delusions and Self Deception: Affective and Motivational Influences on Belief-Formation*, eds. Tim Bayne, Jordi Fernandez, 227-242. Hove, UK: Psychology Press.

Lockie, Robert. 2003. Depth Psychology and Self-Deception. *Philosophical Psychology* 16 (1) (March): 127-48.

Martin, Mike. 1986. *Self-Deception and Morality*. Lawrence, KS: University Press of Kansas.

McLaughlin, Brian P., and Amélie Oksenberg Rorty, eds. 1988. *Perspectives on Self-Deception*. Berkeley, CA: University of California Press.

Mele, Alfred R. 2007. Self-Deception and Hypothesis Testing. In *Cartographies of the Mind: Philosophy and Psychology in Intersection*, eds. Massimo Marraffa, Mario De Caro and Francesco Ferretti, 159-167. Dordrecht: Springer.

———. 2004. *Motivated Irrationality*. The Oxford Handbook of Rationality, ed. Alfred R. Mele. Oxford: Oxford University Press.

———. 2001. *Self-Deception Unmasked*. Princeton Monographs in Philosophy. Princeton, N.J.: Princeton University Press.

———. 1997. Real Self-Deception. *Behavioral and Brain Sciences* 20 (1): 91-102.

———. 1994. Self-Control and Belief. *Philosophical Psychology* 7 (4): 419-35.

———. 1987. *Irrationality : An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press.

———. 1983. Self-Deception. *Philosophical Quarterly* 33 (October): 366-77.

Mele, Alfred R., and Paul Moser. 1997. Intentional Action. In *The Philosophy of Action.*, ed. Alfred R. Mele, 223-255. Oxford: Oxford University Press.

Michel, Christoph, and Albert Newen. 2010. Self-deception as 'pseudo-rational' defense of belief. *Consciousness and Cognition*. Forthcoming .

Moran, Richard. 2004. Anscombe on Practical Knowledge. *Philosophy* 55 (Supp.) : 43-68.

Moran, Richard, and Martin Stone. 2009. Anscombe on Expression of Intention. In *New Essays on the Explanation of Action*, ed. Constantine Sandis, 132-168. Basingstoke: Palgrave Macmillan.

Morris, Errol. The Anosognosic's Dilemma: Something's Wrong But You'll Never Know What It Is. *The New York Times*. 2010. Available from http://opinionator.blogs.nytimes.com/2010/06/23/the-anosognosics-dilemma-somethings-wrong-but-youll-never-know-what-it-is-part-4/.

Nachson, Israel. 2001. Truthfulness, Deception and Self-Deception in Recovering True and False Memories of Child Sexual Abuse *International Review of Victimology* 8 (1): 1-18.

Nisbett, Richard E., and Lee Ross. 1980. *Human Inference : Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, N.J.: Prentice-Hall.

Pears, David. 1991. Self-Deceptive Belief-Formation *Synthese* 89 (3, Belief and Rationality) (Dec.): 393-405.

———. 1984. *Motivated Irrationality*. New York: Oxford University Press.

Potter, Nelson. 2002. Duties to Oneself, Motivational Internalism, and Self-Deception in Kant's Ethics. In *Kant's Metaphysics of Morals: Interpretive Essays*, ed. Mark Timmons, 371-390. Oxford: Oxford University Press.

Quattrone, George, and Amos Tversky. 1984. Causal Versus Diagnostic Contingencies: On Self-Deception and On the Voter's Illusion. *Journal of Personality and Social Psychology* 46 (2): 237-48.

Roland Bénabou, and Jean Tirole. 2002. Self-Confidence and Personal Motivation. *The Quarterly Journal of Economics* 117 (3) (Aug.): 871-915.

Rorty, Amélie Oksenberg. 1994. User-Friendly Self-Deception. *Philosophy* 69 (268) (Apr.): 211-28.

———. 1988. The Deceptive Self: Liars, Layers, And Lairs. In *Perspectives On Self-Deception*, eds. Brian P. McLaughlin, Amélie Oksenberg Rorty, 11-28. Berkeley, CA: University of California Press.

Ross, Stephanie. 1983. Review: [untitled]. *The Philosophical Review* 92 (4) (Oct.): 630-3.

Sahdra, Baljinder. 2003. Self-Deception And Emotional Coherence. *Minds and Machines* 13 (2): 213.

Sartre, Jean-Paul. 1956. *Being and Nothingness*. Trans. Hazel Barnes. New York: Philosophical Library.

Schälike, Julius. 2004. Willensschwäche und Selbsttäuschung. *Deutsche Zeitschrift Für Philosophie* 52 (3): 361-79.

Scott-Kakures, Dion. 2009. Unsettling Questions: Cognitive Dissonance in Self-Deception. *Social Theory and Practice* 35 (1): 73.

———. 2002. At Permanent Risk: Reasoning and Self-Knowledge in Self-Deception *Philosophy and Phenomenological Research* 65 (3) (November): 577-603.

———. 2001. Review: [untitled]: Seeing Through Self-Deception. *Philosophy and Phenomenological Research* 63 (1) (Jul.): 242-5.

———. 1996. Self-Deception and Internal Irrationality. *Philosophy and Phenomenological Research* 56 (1) (Mar.): 31-56.

Shafer-Landau, Russ. 2010. *The Fundamentals of Ethics*. New York: Oxford University Press.

Shapiro, David. 1996. On the Psychology of Self-Deception. *Social Research* 63 (3): 785-800.

Talbott, W. J. 1995. Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research* 55 (1) (Mar.): 27-74.

Taylor, Shelley E. 1989. *Positive illusions: Creative Self-Deception and the Healthy Mind*. New York: Basic Books.

Thagard, Paul, and Baljinder Sahdra. 2003. Self-Deception and Emotional Coherence. *Minds and Machines* 13 (2): 213.

Thompson, Michael. 2008. Naive Action Theory. In *Life and Action*. Vol. 2010, 85-148. Harvard: Harvard University Press.

van Fraasen, Bas. 1988. The Peculiar Effects Of Love And Desire. In *Perspectives on Self-Deception*, eds. Brian P. McLaughlin, Amélie Oksenberg Rorty, 123-156. Berkeley, CA: University of California Press.

Van Leeuwen, D. S. Neil. 2009. Self-Deception Won't Make You Happy. *Social Theory and Practice: An International and Interdisciplinary Journal of Social Philosophy* 35 (1) (January): 107-32.

———. 2008. Finite Rational Self-Deceivers. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 139 (2) (May): 191-208.

———. 2007. The Product of Self-Deception. *Erkenntnis: An International Journal of Analytic Philosophy* 67 (3) (November): 419-37.

———. 2007. The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon. *Philosophical Psychology* 20 (3) (June): 329-48.

Velleman, David. 2000. The Guise of the Good. In *The Possibility of Practical Reason*, 99-122. Oxford: Oxford University Press.

Williams, Bernard. 1973. Deciding to Believe. In *Problems of the Self*, 136-151. Cambridge: Cambridge University Press.

Witschen, Dieter. 2008. Kultivierung des Gewissens - eine Pflicht Gegenüber Sich Selbst: Kantische Reflexionen. *Freiburger Zeitschrift Für Philosophie Und Theologie* 55 (1): 129-41.