COPYRIGHT NOTICE:

Alfred R. Mele: Self-Deception Unmasked

is published by Princeton University Press and copyrighted, © 2001, by Princeton University Press. All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher, except for reading and browsing via the World Wide Web. Users are not permitted to mount this file on any network servers.

For COURSE PACK and other PERMISSIONS, refer to entry on previous page. For more information, send e-mail to permissions@pupress.princeton.edu

1

Introduction: Approaches, Puzzles, Biases, and Agency

"A survey of university professors found that 94% thought they were better at their jobs than their average colleague" (Gilovich 1991, p. 77). Are university professors exceptionally adept at self-deception? Perhaps not. "A survey of one million high school seniors found that . . . *all* students thought they were above average" in their "ability to get along with others . . . and 25% thought they were in the top 1%" (ibid.). One might suspect that the respondents to these surveys were not being entirely sincere in their answers. Then again, how many university professors do you know who do *not* think that they are better at what they do than their average colleague?

Data such as these suggest that we sometimes deceive ourselves. That suggestion raises some interesting questions. How do we deceive ourselves? Why do we deceive ourselves? What is it to deceive oneself? Is self-deception even possible? These questions guide my discussion in this book.

Some theorists understand self-deception as largely isomorphic with stereotypical interpersonal deception. This understanding, which has generated some much discussed puzzles or "paradoxes," guides influential work on self-deception not

only in philosophy but also in psychology, psychiatry, and biology.[1] In the course of resolving the major puzzles, I argue that the attempt to understand self-deception on the model of stereotypical interpersonal deception is fundamentally misguided. The position on self-deception defended here is *deflationary*. If I am right, self-deception is neither irresolvably paradoxical nor mysterious, and it is explicable without the assistance of mental exotica. Although a theorist whose interest in self-deception is restricted to the outer limits of logical or conceptual possibility might view this as draining the topic of conceptual intrigue, the main source of broader, enduring interest in self-deception is a concern to understand and explain the behavior of real human beings.

1. Preview

Self-deception apparently occurs in two quite different forms, "straight" and "twisted." Straight cases of self-deception have received pride of place in philosophical and empirical work. In these cases, people are self-deceived in believing something that they want to be true—for example, that they are not seriously ill, that their children are not experimenting with drugs, or that a loved one is innocent of a criminal charge. In twisted cases, people are self-deceived in believing something that they want to be false (and do not also want to be true). For example, an insecure, jealous husband may believe that his wife is having an affair despite his possessing only relatively flimsy evidence for that proposition and despite his not wanting it to be the case that she is so engaged.[2] If some self-deception is twisted in this sense, at least one relatively common claim about self-deception is false—the claim that *S*'s being self-deceived that *p* requires *S*'s desiring that *p*.[3] Furthermore, twisted self-deception apparently threatens even the more modest claim that all

self-deception is motivated or has a motivated component.[4] Although the most obvious antonym of "straight" is "bent," I prefer "twisted" here for stylistic reasons. I am not using the term pejoratively and I do not regard twisted self-deception as essentially pathological.)

In Chapters 2 and 3, I offer an account of the nature and etiology of garden-variety straight self-deception and resolve some familiar puzzles about self-deception. In Chapter 4, I review and reject attempted empirical demonstrations of a "strict" kind of self-deception in which the self-deceiver believes a proposition, *p*, while also believing its negation, ~*p*. In Chapter 5, I develop a pair of approaches to explaining twisted self-deception—a motivation-centered approach and a hybrid approach featuring both motivation and emotion—in order to display our resources for exploring and explaining twisted self-deception and to show that promising approaches are consistent with my position on straight self-deception.

2. Three Approaches to Characterizing Self-Deception and a Pair of Puzzles

In defining self-deception, three common approaches may be distinguished: *lexical*, in which a theorist starts with a definition of "deceive" or "deception," using the dictionary or common usage as a guide, and then employs it as a model for defining self-deception; *example-based*, in which one scrutinizes representative examples of self-deception and attempts to identify their essential common features; and *theory-guided*, in which the search for a definition is guided by commonsense theory about the etiology and nature of self-deception. Hybrids of these approaches are also common.

The lexical approach may seem safest. Practitioners of the example-based approach run the risk of considering too narrow

a range of cases. The theory-guided approach, in its typical manifestations, relies on commonsense explanatory hypotheses that may be misguided: even if ordinary folks are good at identifying hypothetical cases of self-deception, they may be quite unreliable at diagnosing what happens in them. In its most pristine versions, the lexical approach relies primarily on a dictionary definition of "deceive." And what could be a better source of definitions than the dictionary?

Matters are not so simple, however. There are weaker and stronger senses of "deceive" both in the dictionary and in common parlance. Lexicalists need a sense of "deceive" that is appropriate to self-deception. On what basis are they to identify that sense? Must they eventually turn to representative examples of self-deception or to commonsense theories about what happens in instances of self-deception?

The lexical approach is favored by theorists who deny that self-deception is possible (e.g., Gergen 1985; Haight 1980; Kipp 1980). A pair of lexical assumptions is common:

> 1. By definition, person *A* deceives person *B* (where *B* may or may not be the same person as *A*) into believing that *p* only if *A* knows, or at least believes truly, that ~*p* and causes *B* to believe that *p*.
> 2. By definition, deceiving is an intentional activity: nonintentional deceiving is conceptually impossible.

Each assumption is associated with a familiar puzzle about self-deception.

If assumption 1 is true, then deceiving oneself into believing that *p* requires that one knows, or at least believes truly, that ~*p* and causes oneself to believe that *p*. At the very least, one starts out believing that ~*p* and then somehow gets oneself to believe that *p*. Some theorists take this to entail that, at some time, self-deceivers both believe that *p* and believe that ~*p* (e.g., Kipp 1980, p. 309). And, it is claimed, this is not a possible state of mind: the very nature of belief precludes one's simultaneously

believing that *p* is true and believing that *p* is false.[5] Thus we have a *static* puzzle about self-deception: self-deception, according to the view at issue, requires being in an impossible *state of mind*.

In fact, assumption 1 does not entail that in all instances of deceiving, there is some time at which the deceiver believes that ~*p* and the deceived person believes that *p*. In some cases of interpersonal deception, *A* has ceased believing that ~*p* by the time he causes *B* to believe that *p*. Imagine that the vehicle for *A*'s attempted deception is a letter. In his letter, *A* attempts to deceive *B* into believing that *p* by lying to him: *p* is false and his assertion of *p* in the letter is a lie. When he sends the letter, *A* is confident that ~*p*, but he comes to believe that *p* by the time *B* receives the letter. If *A*'s lie is successful, *A* deceives *B* into believing that *p* in a way that provides confirmation for assumption 1. But there is no time at which *A* believes that ~*p* and *B* believes that *p* (see Sorensen 1985).

A theorist inclined to believe that there is a basis in "the concept of deception" for the claim that self-deceivers simultaneously believe that *p* and believe that ~*p* need not be undone by the preceding observation. It may well be true that in *stereotypical* cases of interpersonal deceiving there is some time at which *A* believes that ~*p* and *B* believes that *p*. And it is open to a theorist to contend that self-deception is properly understood only on the model of stereotypical interpersonal deception.

The claim that self-deception must be understood on the model just mentioned produces a further puzzle about the state of self-deception. In stereotypical cases of interpersonal deceiving, there is a time at which the deceiver does *not* have a belief that *p* and the deceived person does have a belief that *p*. If self-deception is strictly analogous to stereotypical interpersonal deception, there is a time at which the self-deceiver both has a belief that *p* and does not have a belief that *p*—a perplexing condition, indeed.[6]

Assumption 2 generates a *dynamic* puzzle, a puzzle about the dynamics of self-deception. On the one hand, it is hard to imagine how one person can deceive another into believing that *p* if the latter person knows exactly what the former is up to, and it is difficult to see how the trick can be any easier when the intending deceiver and the intended victim are the same person. On the other, deception normally is facilitated by the deceiver's having and intentionally executing a deceptive strategy. If, to avoid thwarting one's own efforts at self-deception, one must not intentionally execute any strategy for deceiving oneself, how can one succeed? The challenge is to explain how self-deception in general is a psychologically possible process. If self-deceivers intentionally deceive themselves, one wonders what prevents the guiding intention from undermining its own effective functioning. And if self-deception is not intentional, what motivates and directs processes of self-deception?[7]

A theorist who believes that self-deception is a genuine phenomenon may attempt to solve the puzzles while leaving assumptions 1 and 2 unchallenged. An alternative tack is to undermine these assumptions and to display the relevance of their falsity to a proper understanding of self-deception. That is the line I pursue.

Stereotypical instances of deceiving someone else into believing that *p* are instances of intentional deceiving in which the deceiver knows or believes truly that ~*p*. Recast as claims specifically about *stereotypical* interpersonal deceiving, assumptions 1 and 2 would be acceptable. But in their present formulations the assumptions are false. In a standard use of "deceived" in the passive voice, we properly say such things as "Unless I am deceived, I left my keys in my car." Here "deceived" means "mistaken." There is a corresponding use of "deceive" in the active voice. In this use, to deceive is "to cause to believe what is false," according to the *Oxford English Dictionary*. Obviously, one can intentionally or unintentionally cause someone to believe what is false; and one can cause someone to acquire the

false belief that *p* even though one does not oneself believe that ~*p*. Yesterday, mistakenly believing that my daughter's schoolbooks were on my desk, I told her they were there. In so doing, I caused her to believe a falsehood. I deceived her, in the sense identified; but I did not do so intentionally, nor did I cause her to believe something I disbelieved.

The point just made has little significance for self-deception, *if* paradigmatic instances of self-deception have the structure of stereotypical instances of interpersonal deception. But do they? Stock examples of self-deception, both in popular thought and in the literature, feature people who falsely believe—in the face of strong evidence to the contrary—that their spouses are not having affairs, or that their children are not using illicit drugs, or that they themselves are not seriously ill. Is it a plausible diagnosis of what happens in such cases that these people start by knowing or believing the truth, *p*, and intentionally cause themselves to believe that ~*p*? If, in our search for a definition of self-deception, we are guided partly by these stock examples, we may deem it an open question whether self-deception requires intentionally deceiving oneself, getting oneself to believe something one earlier knew or believed to be false, simultaneously possessing conflicting beliefs, and the like. If, instead, our search is driven by a presumption that nothing counts as self-deception unless it has the same structure as stereotypical interpersonal deception, the question is closed at the outset.

Theorists who accept lexical assumptions 1 and 2 may proceed in either of two ways when confronting cases that most people would count as clear instances of self-deception. They may suppose that many such cases are not properly so counted because they fail to satisfy one or both of the assumptions. Alternatively, they may suppose that all or most cases that would generally be deemed clear instances of self-deception do, in fact, satisfy the lexical assumptions, even if they may seem not to. On either alternative, self-deception as a whole is made to

seem puzzling. And on the second alternative, as I argue, puzzles are generated in cases that are describable and explicable in quite unpuzzling ways.

Compare the question whether self-deception is properly understood on the model of stereotypical interpersonal deception with the question whether addiction is properly understood on the model of disease. The current folk conception of addiction seemingly treats addictions as being, by definition, diseases. The disease model of addiction, however, has been forcefully attacked (see, e.g., Peele 1989). The issue is essentially about explanation, not about alleged conceptual truths. How is the characteristic behavior of people typically counted as addicts best explained? Is the disease model of addiction explanatorily more accurate or fruitful than its competitors? Self-deception, like addiction, is an explanatory concept. We postulate self-deception in particular cases to explain data: for example, the fact that there are excellent grounds for holding that *S* believes that *p* despite its being the case that evidence *S* possesses makes it quite likely that ~*p*. And we should ask how self-deception is likely to be constituted—what it is likely to be—if it does help to explain the relevant data. Should we discover that the data explained by self-deception are *not* explained by a phenomenon involving the simultaneous possession of beliefs whose contents are mutually contradictory or intentional acts of deception directed at oneself, self-deception would not disappear from our conceptual map—any more than addiction would disappear should we learn that addictions are not diseases.

An announcement about belief is in order before I move forward. In the literature on self-deception, belief rather than degree of belief usually is the operative notion. I follow suit in this book, partly to avoid unnecessary complexities. Those who prefer to think in terms of degree of belief should read such expressions as "*S* believes that *p*" as shorthand for "*S* believes that *p* to a degree greater than 0.5 (on a scale from 0 to 1)."[8]

3. Motivationally Biased Belief and Agency

That there are motivationally biased beliefs is difficult to deny. In a passage from which I quoted at the beginning of this chapter, Thomas Gilovich reports:

> A survey of one million high school seniors found that 70% thought they were above average in leadership ability, and only 2% thought they were below average. In terms of ability to get along with others, *all* students thought they were above average, 60% thought they were in the top 10%, and 25% thought they were in the top 1%! . . . A survey of university professors found that 94% thought they were better at their jobs than their average colleague. (1991, p. 77)

If we assume the sincerity of the people surveyed, a likely hypothesis is that motivation had a hand in producing many of the beliefs reported. The aggregated self-assessments are radically out of line with the facts (e.g., only 1 percent can be in the top 1 percent), and the qualities asked about are desirable ones. We may have a tendency to believe propositions that we want to be true even when an impartial investigation of readily available data would indicate that they are probably false. A plausible hypothesis about that tendency is that our desiring something to be true sometimes exerts a biasing influence on what we believe. And there is evidence that our beliefs about our own traits "become more biased when the trait is highly desirable or undesirable" (Brown and Dutton 1995, p. 1290).

Ziva Kunda ably defends the view that motivation can influence "the generation and evaluation of hypotheses, of inference rules, and of evidence," and that motivationally "biased memory search will result in the formation of additional biased beliefs and theories" that cohere with "desired conclusions" (1990, p. 483). In an especially persuasive study, undergraduate

subjects (seventy-five women and eighty-six men) read an article alleging that "women were endangered by caffeine and were strongly advised to avoid caffeine in any form"; that the major danger was fibrocystic disease, "associated in its advanced stages with breast cancer"; and that "caffeine induced the disease by increasing the concentration of a substance called cAMP in the breast" (Kunda 1987, p. 642). (Because the article did not personally threaten men, they were used as a control group.) Subjects were then asked to indicate, among other things, "how convinced they were of the connection between caffeine and fibrocystic disease and of the connection between caffeine and . . . cAMP on a 6-point scale" (pp. 643–44). In the female group, "heavy consumers" of caffeine were significantly less convinced of the connections than were "low consumers." The males were considerably more convinced than the female "heavy consumers"; and there was a much smaller difference in conviction between "heavy" and "low" male caffeine consumers (the heavy consumers were slightly *more* convinced of the connections).

Because all subjects were exposed to the same information and arguably only the female "heavy consumers" were personally threatened by it, a plausible hypothesis is that their lower level of conviction is motivated in some way by a desire that their coffee drinking has not significantly endangered their health (cf. Kunda 1987, p. 644). Indeed, in a study in which the reported hazards of caffeine use were relatively modest, "female heavy consumers were no less convinced by the evidence than were female low consumers" (p. 644). Along with the lesser threat, there is less motivation for skepticism about the evidence.

How do the female heavy consumers come to be less convinced than the others? One testable possibility is that because they find the "connections" at issue personally threatening, these women (or some of them) are motivated to take a hypercritical stance toward the article, looking much harder than other subjects for reasons to be skeptical about its merits (cf.

Kunda 1990, p. 495; Liberman and Chaiken 1992). Another is that, owing to the threatening nature of the article, they (or some of them) read it *less* carefully than the others do, thereby enabling themselves to be less impressed by it.[9] In either case, must we suppose that the women intend to deceive themselves, or intend to bring it about that they hold certain beliefs, or start by finding the article convincing and then try to get themselves to find it less convincing? Or can motivation issue in biased beliefs without the assistance of such intentions or efforts?

Consider the following two bold theses about motivationally biased beliefs.

> 1. The agency view: all motivationally biased beliefs are intentionally produced or protected. In every instance of motivationally biased belief that *p*, we try to bring it about that we acquire or retain the belief that *p*, or at least try to make it easier for ourselves to acquire or retain the belief.
> 2. The antiagency view: no motivationally biased beliefs are intentionally produced or protected. In no instance of motivationally biased belief that *p* does one try to bring it about that one acquires or retains the belief or try to make it easier for oneself to acquire or retain the belief.

One suspects that the truth lies somewhere between these poles. But which of the two theses is likely to be closer to the truth?

One problem for the agency view is central to the dynamic puzzle about self-deception. The attempts to which the view appeals threaten to undermine themselves. If I am trying to bring it about that I believe that I am a good driver—not by improving my driving skills, but perhaps by ignoring or downplaying evidence that I am an inferior driver while searching for evidence of my having superior driving skills—won't I see that the "grounds" for belief that I arrive at in this way are

illegitimate? And won't I therefore find myself still lacking the belief that I am a good driver. A predictable reply is that the "tryings" or efforts to which the agency view appeals are not conscious efforts and therefore need not stand in the way of their own success in the way just envisioned. Whether, and to what extent, we should postulate unconscious tryings in attempting to explain motivationally biased belief depends on what the alternatives are.

The main problem for the antiagency view is also linked to the dynamic puzzle about self-deception. Apparently, we encounter difficulties in trying to understand how motivationally biased beliefs—or many such beliefs—can arise, if not through efforts of the kind the agency view postulates. How, for example, can my wanting it to be the case that I am a good driver motivate me to believe that I am a good driver except by motivating me to try to bring it about that I believe this or by motivating me to try to make it easier for myself to believe this?[10] At the very least, the antiagency view is faced with a clear challenge: to provide an alternative account of the mechanism(s) by which desires lead to motivationally biased beliefs. I take up this challenge in Chapters 2 and 3, in developing a position on the nature and etiology of garden-variety straight self-deception, and I return to it in Chapter 4, in rebutting an alleged empirical demonstration of "strict" self-deception.

Ideally, in exploring the relative merits of the agency and antiagency views, one would start with uncontroversial analyses of *intentional action* and *trying*. Paul Moser and I have offered an analysis of intentional action (Mele and Moser 1994), and Frederick Adams and I have offered an account of trying (Adams and Mele 1992). If I were to deem these offerings uncontroversial, however, the hypothesis that I am merely self-deceived would be quite generous. Fortunately, for the purposes of this book, full-blown analyses of these notions are not required. But some conceptual spade work is in order.

The question how much control an agent must have over an outcome for that outcome to count as *intentionally* produced has elicited strikingly opposed intuitions. According to Christopher Peacocke, it is "undisputed" that an agent who makes a successful attempt "to hit a croquet ball through a distant hoop" *intentionally* hits the ball through the hoop (1985, p. 69). But Brian O'Shaughnessy maintains that a novice who similarly succeeds in hitting the bull's-eye on a dart board does not intentionally hit the bull's-eye (1980, 2:325; cf. Harman 1986, p. 92). This conceptual issue can be skirted, for the purposes of this book, by focusing on whether people who acquire motivationally biased beliefs that *p try* to bring it about that they acquire beliefs that *p*, or try to make it easier for themselves to acquire these beliefs. If they do try to do this, one need not worry about whether the success of their attempts owes too much to luck, or to factors beyond the agents' control, for it to be true that they *intentionally* brought it about that they believed that *p*. (Trying to *A*, as I understand it, does not require making a *special* effort to *A*. When I typed the word "special" a moment ago, I was trying to do that, even though I encountered no special resistance and made no remarkable effort to type it.)

Furthermore, if they do *not* try to do this, there is, I believe, no acceptable sense of "intentionally" in which they intentionally bring it about that they believe that *p*. Unfortunately, here one confronts another controversy in the philosophy of action. Some philosophers contend that an agent who tries to do *A*, recognizing that her doing *B* is a likely consequence of her doing *A*, may properly be said to do *B* intentionally (if she does *B*), even if she does not try to do *B* and is in no way attracted to doing *B* (e.g., as a means or as an end), and even if she prefers that her doing *A* not have her doing *B* as a side effect (Bratman 1987, chs. 8–10; Harman 1976). Others reject this idea, contending, roughly, that aside from tryings themselves, we intentionally do only what we try to do (Adams 1986; McCann

1986b, 1991; Mele and Moser 1994; O'Shaughnessy 1980). Steven Sverdlik and I have criticized the grounds for the former view (Mele and Sverdlik 1996), and I do not reopen the debate here. For present purposes, the crucial question is whether motivated beliefs that one is self-deceived in holding are (necessarily, always, or ordinarily) beliefs that one *tries* to bring about or promote. Theorists who favor an affirmative answer often deem the trying involved—or the associated intentions—to be unconscious (Bermudez 1997; Martin 1997; Talbott 1995, 1997), and I assume, accordingly, that unconscious tryings and intentions are possible.

Intentionally deceiving oneself is unproblematically possible. It is worth noting, however, that the unproblematic cases are remote from garden-variety self-deception. Here is an illustration. Ike, a forgetful prankster skilled at imitating others' handwriting, has intentionally deceived friends by secretly making false entries in their diaries. Ike has just decided to deceive himself by making a false entry in his own diary. Cognizant of his forgetfulness, he writes under today's date, "I was particularly brilliant in class today," counting on eventually forgetting that what he wrote is false. Weeks later, when reviewing his diary, Ike reads this sentence and acquires the belief that he was brilliant in class on the specified day. If Ike intentionally deceived others by making false entries in their diaries, what is to prevent us from justifiably holding that he intentionally deceived himself in the imagined case? He intended to bring it about that he would believe that *p*, which he knew at the time to be false; and he executed that intention without a hitch, causing himself to believe, eventually, that *p*. Again, to deceive, on one standard definition, is to cause to believe what is false; and Ike's causing himself to believe the relevant falsehood is no less intentional than his causing his friends to believe falsehoods (by doctoring their diaries).[11]

Ike's case undoubtedly strikes readers as markedly dissimilar to garden-variety examples of self-deception—for instance, the

case of the woman who falsely believes that her child is not using drugs (or that she is healthy or that her husband is not having an affair), in the face of strong evidence to the contrary. Why is that? The most obvious difference between Ike's case and garden-variety examples of self-deception lies in the straightforwardly intentional nature of Ike's project. Ike consciously sets out to deceive himself and he intentionally and consciously executes his plan for so doing; ordinary self-deceivers behave quite differently.[12]

This suggests that in attempting to construct hypothetical cases that are, at once, paradigmatic cases of self-deception and cases of agents intentionally deceiving themselves, one should imagine that the agents' intentions to deceive themselves are somehow hidden from them. I do not wish to claim that "hidden intentions" are impossible. Our ordinary concept of intention may leave room, for example, for "Freudian" intentions, hidden in some mental partition. And if there is conceptual space for hidden intentions that play a role in the etiology of behavior, there is conceptual space for hidden intentions to deceive ourselves, intentions that may influence our treatment of data. As I see it, the claim is *unwarranted*, *not* incoherent, that intentions to deceive ourselves, or intentions to produce or sustain certain beliefs in ourselves, or corresponding attempts—normally, intentions or attempts hidden from us—are at work in ordinary self-deception.[13] Without denying that "hidden intention" or "hidden attempt" cases of self-deception are possible, a theorist should ask what evidence there may be (in the real world) that intentions or attempts to deceive oneself, or to make it easier for oneself to believe something, are at work in garden-variety self-deception. Are there data that can *only*—or *best*—be explained on the hypothesis that such intentions or attempts are operative in such self-deception? The answer that I defend in subsequent chapters is *no*.

Distinguishing activities of the following three kinds will prove useful. Regarding cognitive activities that contribute to

motivationally biased belief, there are significant differences among (1) *unintentional* activities (e.g., unintentionally focusing on data of a certain kind), (2) *intentional* activities (e.g., intentionally focusing on data of a certain kind), and (3) intentional activities engaged in as part of an *attempt* to deceive oneself, or to cause oneself to believe something, or to make it easier for oneself to believe something (e.g., intentionally focusing on data of a certain kind as part of an attempt to deceive oneself into believing that *p*). Many skeptical worries about the reality of self-deception are motivated partly by the assumption that activity of the third kind is characteristic of self-deception.

An important difference between the second and third kinds of activity merits emphasis. Imagine a twelve-year-old, Beth, whose father died some months ago. Beth may find it comforting to reflect on pleasant memories of playing happily with her father, to look at family photographs of such scenes, and the like. Similarly, she may find it unpleasant to reflect on memories of her father leaving her behind to play ball with her brothers, as he frequently did. From time to time, she may intentionally focus her attention on the pleasant memories, intentionally linger over the pictures, and intentionally turn her attention away from memories of being left behind and from pictures of her father playing only with her brothers. As a consequence of such intentional activities, she may acquire a false, unwarranted belief that her father cared more deeply for her than for anyone else. Although her intentional cognitive activities may be explained, in part, by the motivational attractiveness of the hypothesis that he loved her most, those activities need not also be explained by a desire—much less an intention or an attempt—to deceive herself into believing this hypothesis, or to cause herself to believe this, or to make it easier for herself to believe this. Intentional cognitive activities that contribute even in a relatively straightforward way to motivationally biased, false, unwarranted belief need not be guided by an intention of

any of the kinds just mentioned, nor need they involve associated attempts to manipulate what one believes. Beth's activities are explicable on the hypothesis that she was seeking pleasant experiences and avoiding painful ones without in any way trying to influence what she believed. Whether a case like the present one is plausibly counted as an instance of self-deception remains to be seen.

Obviously, an agent's doing something that he is trying to do can have a result that he does not try to produce. Intending to turn on a light in an unfamiliar kitchen, Al tries to flip the switch on his left, and he succeeds in flipping it. As it happens, that switch is wired to the garbage disposal. So Al turns on the garbage disposal, but he does not try to do that. Similarly, Beth tries to focus her attention on certain memories and photographs and tries to avoid focusing it on certain other things, and she succeeds in this. Perhaps, in doing these things, she is also trying to comfort herself. Beth's cognitive activities result in her believing that her father loved her most. But, clearly, these points do not entail that Beth is trying to produce this belief or trying to make it easier for herself to acquire this belief—any more than similar points about Al entail that he is trying to activate the garbage disposal.

Another illustration of the difference between the second and third kinds of activity may prove useful. Donald Gorassini has suggested that an intentional form of self-deception is quite common (1997, p. 116). Described in a theory-neutral way, what Gorassini has in mind are cases in which a person who lacks a certain quality—for example, kindness—but is desirous of its being true that he has that quality is motivated to act *as if* he has it and then infers from his behavior that he does have it. I discussed cases of this kind previously under the rubric "acting as if" (Mele 1987a, pp. 151–58). One of the points I made is that an agent's motivation to act as if *p* may have sources of various kinds. Here are two examples. Ann believes that she can

cultivate the trait of kindness in herself by acting as if she were kind; so, because she wants to become kind, she decides to embark on a program of acting as if she were kind, and she acts accordingly. Because Bob would like to be a generous person, he finds pleasure in actions of his that are associated with the trait; consequently, Bob has hedonic motivation to act as if he were generous, and he sometimes acts accordingly. Unlike Ann, Bob is not trying to inculcate the desired trait in himself.

There is considerable evidence that we often make inferences about our qualities on the basis of our own behavior (see, e.g., Bem 1972). It is easy to imagine that, after some time, Ann and Bob infer, largely from their relevant behavior, that they have the desired trait, even though they in fact lack it. However, from the facts that these agents want it to be true that *p*, intentionally act as if *p* owing significantly to their wanting *p* to be true, and come to believe that *p* largely as a consequence of that intentional behavior, it does not follow that they were trying to deceive themselves into believing that *p* or trying to make it easier for themselves to believe that *p*. Ann may simply have been trying to make herself kind and Bob may merely have been seeking the pleasure that acts associated with generosity give him.

A related point may be made about cases in which one's desire that *p* and intentional behavior that it motivates lead to biased beliefs about one's traits via a route that has a major social component. An older boy who is strongly desirous of its being true that he is a natural leader but who lacks the admiration of his peers may find the company of younger, impressionable teenagers considerably more pleasant. His hedonically motivated choice of younger companions may result in selective exposure to data supportive of the hypothesis that he is a natural leader; the younger teenagers might worship him. This choice and the social feedback it helps generate may contribute significantly to his acquiring an unwarranted, biased belief about his leadership ability. But to explain what happens in such a

case, there is no need to suppose that the boy was trying to get himself to believe that he was a natural leader, or trying to make it easier for himself to believe this.

The following remarks by David Pears and Donald Davidson on the self-deceptive acquisition of a motivationally biased belief are concise expressions of two different "agency" views of the phenomenon:

> [There is a] sub-system . . . built around the nucleus of the wish for the irrational belief and it is organized like a person. Although it is a separate centre of agency within the whole person, it is, from its own point of view, entirely rational. It wants the main system to form the irrational belief and it is aware that it will not form it, if the cautionary belief [i.e., the belief that it would be irrational to form the desired belief] is allowed to intervene. So with perfect rationality it stops its intervention. (Pears 1984, p. 87)

> His practical reasoning is straightforward. Other things being equal, it is better to avoid pain; believing he will fail the exam is painful; therefore (other things being equal) it is better to avoid believing he will fail the exam. Since it is a condition of his problem that he take the exam, this means it would be better to believe he will pass. He does things to promote this belief. (Davidson 1985, pp. 145–46)

Both views rest largely on the thought that the only way, or the best way, to account for certain data is to hold that the person, or some center of agency within the person, tries to bring it about that the person, or some "system" in the person, holds a certain belief. In subsequent chapters, I argue that we can account for the pertinent data in more plausible and less problematic ways.

Consider a case of self-deception similar to the one Davidson diagnoses in the passage just quoted. Carlos "has good reason to believe" that he will fail his driver's test (p. 145). "He has

failed the test twice before and his instructor has said discouraging things. On the other hand, he knows the examiner personally, and he has faith in his own charm" (pp. 145–46). "The thought of failing the test once again is painful to Carlos (in fact the thought of failing anything is particularly galling to Carlos)." Suppose that the overwhelming majority of Carlos's impartial cognitive peers presented with his evidence would believe that Carlos will fail the test and that none of them would believe that Carlos will pass it. (Perhaps some peers with particularly high standards for belief would withhold belief.) Even so, in the face of the evidence to the contrary, Carlos believes that he will pass. Predictably, he fails.

If lexical assumption 1 about deception were true (see sec. 2), then, on the assumption that Carlos is self-deceived in believing that he will pass the test, he believed at some time that he would fail the test. In accommodating the data offered in my description of the case, however, there is no evident need to suppose that Carlos had this true belief. Perhaps his self-deception is such that not only does he acquire the belief that he will pass the test, but he never acquires the belief that he will fail. In fact, at least at first sight, it seems that this is true of much self-deception. Seemingly, at least some parents who are self-deceived in believing that their children have never experimented with drugs and some people who are self-deceived in believing that their spouses have not had affairs have at no point believed that these things have happened. Owing to self-deception, they have not come to believe the truth, and perhaps they never will.

That having been said, it does seem that there are cases in which a person who once believed an unpleasant truth, p, later is self-deceived in believing that $\sim p$. For example, a mother who once believed that her son was using drugs subsequently comes to believe that he has never used drugs and is self-deceived in so believing. Does a change of mind of this sort *require* an exercise of agency of the kind postulated by Pears or Davidson?

Is such a change of mind *most plausibly explained*, at least, on the hypothesis that an exercise of agency of one of these kinds occurred? A theorist who attends to the stark descriptions Pears and Davidson offer of the place of agency in self-deception should at least wonder whether things are in fact so straightforward.

It is often supposed that, as one philosopher has put it, (1) "desires have no explanatory force without associated beliefs" that identify means, or apparent means, to the desires' satisfaction and (2) this is part of "the very logic of belief-desire explanation" (Foss 1997, p. 112). Setting aside intentional *A*-ings that are motivated by intrinsic desires to *A* (i.e., desires that treat one's *A*-ing as an end), claim 1 may be part of the logic of belief-desire explanation of *intentional action*.[14] But the claim does not fare well in the sphere or motivationally biased belief.

Recall the "survey of one million high school seniors" that found, among other things, that "25% thought they were in the top 1%" in ability to get along with others (Gilovich 1991, p. 77). The figures are striking, and a likely hypothesis about them includes the idea that desires that *p* can contribute to biased beliefs that *p*. If claim 1 were true, a student's wanting it to be the case that she has superior ability to get along with others would help to explain her believing that she is superior in this area only in conjunction with some instrumental belief that links her believing that she is superior in this area to the satisfaction of her desire to be superior. But one searches in vain for instrumental beliefs that would both turn the trick and be plausibly widely attributed to high school seniors. Perhaps believing that one has a superior ability to get along with others can help to bring it about that one is in fact superior in this sphere, and some high school students might believe that this is so. But it is highly unlikely that most people who have a motivationally biased belief that they have a superior ability to get along with others have this belief, in part, *because* they want it

to be true that they are superior in this area *and* believe that believing that they are superior can make it so. And no other instrumental belief looks more promising.

Should we infer, then, that wanting it to be the case that one has a superior ability to get along with others plays a role in explaining only relatively few instances of false and unwarranted belief that one is superior in this area? Not at all. There is powerful empirical evidence, some of which is reviewed in Chapter 2, that desiring that *p* makes a broad causal contribution to the acquisition and retention of unwarranted beliefs that *p*. Desires that do this properly enter into causal *explanations* of the pertinent biased beliefs. It is a mistake to assume that the role characteristic of desires in explaining intentional actions is the only explanatory role desires can have.

If Pears or Davidson is right about a case like the mother's or Carlos's, presumably similar exercises of agency are at work in an enormous number of high school students who believe that, regarding ability to get along with others, they are "in the top 1%" and in a great many university professors who believe that they are better at what they do than their average colleague. Perhaps self-deception is very common, but the same is unlikely to be true of intentional self-manipulation of the kind Pears or Davidson describes. Theorists inclined to agree with claims 1 and 2 about the explanatory force of desires will be inclined toward some version of the agency view of motivationally biased belief and self-deception. As I will argue, however, desires contribute to the production of motivationally biased beliefs, including beliefs that one is self-deceived in holding, in a variety of relatively well understood ways that fit the anti-agency model.