

CHAPTER 13

MOTIVATED IRRATIONALITY

ALFRED R. MELE

THE literature on motivated irrationality has two primary foci: action and belief. In the former sphere, akratic action—action exhibiting so-called weakness of will or deficient self-control—has received pride of place. Philosophical work on motivated irrational belief includes, but is not limited to, work on self-deception. The primary topics of this chapter are akratic action and motivationally biased belief.

1. AKRASIA, SELF-CONTROL, AND STRICT AKRATIC ACTION

The classical Greek term *akrasia* is formed from the alpha privative and *kratos*—strength or power. The pertinent power is the power to control oneself. Hence, *akrasia* is deficient self-control. Self-control, in this sense, is, very roughly, a robust capacity to see to it that one acts as one judges best in the face of actual or anticipated competing motivation.¹ The trait may be either regional or global (Rorty 1980a). A scholar who exhibits remarkable self-control in adhering to the

demanding work schedule that she deems best for herself may be akratic about smoking. She is self-controlled in one region of her life and akratic in another. Self-control also comes in degrees: some self-controlled individuals are more self-controlled than others. People with global self-control—self-control in all regions of their lives—would be particularly remarkable, if, in every region, their self-control considerably exceeded that of most people.

In Plato's *Protagoras*, Socrates says that the common view about akratic action is that "many people who know what it is best to do are not willing to do it, though it is in their power, but do something else" (Plato 1953, 352d). Here he raises (among other issues) the central question in subsequent philosophical discussion of *akrasia*: Is it possible to perform uncompelled intentional actions that, as one recognizes, are contrary to what one judges best, the judgment being made from the perspective of one's values, principles, desires, and beliefs? More briefly, is *strict* akratic action possible (Mele 1987, 7)? Relevant judgments include judgments in which "best" is relativized to options envisioned by the agent at the time. In strict akratic action, an agent need not judge that a course of action *A* is the best of all possible courses of action open to her then. She may judge that *A* is better than the alternatives she has envisioned—that it is the best of the envisioned options. If, nevertheless, in the absence of compulsion, she does not *A* and intentionally pursues one of the envisioned alternatives, she acts akratically. It is a truism that a perfectly self-controlled agent would never act akratically. So akratic action, if it is possible, exhibits at least imperfect self-control.²

The judgment against which an agent acts in strict akratic action is what I call a *decisive* judgment. An agent's judgment that *A* is the best of his envisioned options at a given time is a decisive judgment, in my sense, if and only if it *settles* in the agent's mind the question of which member of the set is best (from the perspective of his own desires, beliefs, etc.)—and best not just in some respect or other (e.g., financially), but without qualification.³ Ann judges that *A*-ing would be morally (or aesthetically, or economically) better than *B*-ing and yet, in the absence of compulsion, she intentionally *B*-s rather than *A*-s. In *B*-ing, Ann need not be acting akratically, for she may also judge, for example, that *all things considered*, *B*-ing would be better than *A*-ing.

A feature of paradigmatic strict akratic actions that typically is taken for granted and rarely made explicit is that the judgments with which they conflict are *rationally* formed. In virtue of their clashing with the agent's rationally formed decisive judgment, such actions are subjectively irrational (to some degree, if not without qualification). There is a failure of coherence in the agent of a kind directly relevant to assessments of the agent's rationality.⁴

The occurrence of strict akratic actions seems to be an unfortunate fact of life. Unlike many such (apparent) facts, however, this one has attracted considerable philosophical attention for nearly two and a half millennia. A major source of the interest is obvious: strict akratic action raises difficult questions about the

connection between evaluative judgment and action, a connection of paramount importance for any theory of the explanation of intentional behavior that accords evaluative judgments an explanatory role.

Matters are complicated by our having—both in various theoretical approaches to understanding action and in ordinary thought—a pair of perspectives on the explanation of intentional action, a *motivational* and an *intellectual* one (Mele 1995, 16–19; Pettit and Smith 1993). Central to the motivational perspective is the idea that what agents do when they act intentionally depends on where their strongest motivation lies then.⁵ This perspective is taken on *all* intentional action, independently of the species to which the agents belong. If cats, dogs, and human beings act intentionally, the motivational perspective has all three species in its sights. The intellectual perspective applies only to intellectual beings. Identifying minimally sufficient conditions for membership in the class of intellectual beings is an ambitious task, but it is clear that the work of practical intellect, as it is normally conceived, includes weighing options and making judgments about what it is best, better, or “good enough” to do. Central to the intellectual perspective is the idea that such judgments play a significant role in explaining some intentional actions.

Many philosophers seek to combine these two perspectives into one in the domain of intentional human action. One tack is to insist that, in intellectual beings, motivational strength and evaluative judgment always are mutually aligned. Socrates seemingly advances this view in connection with his thesis that people never knowingly do wrong (Plato 1953, 352b–358d). Theorists who take this tack have various options. For example, they can hold that judgment causally determines motivational strength, that motivational strength causally determines judgment, or that judgment and motivational strength have a common cause. They can also try to get by without causation, seeking purely conceptual grounds for the alignment thesis.

The apparent occurrence of strict akratic actions is a problem for this general tack. The motivational perspective is well suited to akratic action: when acting akratically, one presumably does what one is most strongly motivated to do then. But the intellectual perspective is threatened; more precisely, certain interpretations of, or theses about, that perspective are challenged. In threatening the intellectual perspective while leaving the motivational perspective unchallenged, akratic action poses apparent difficulties for the project of combining the two perspectives into a unified outlook on the explanation of intentional human action. That is a primary source of perennial philosophical interest in akratic action.

There is much to recommend the motivational and intellectual perspectives, and a plausible combination is theoretically desirable. To some theorists, the threat that strict akratic action poses to a unified, motivational/intellectual perspective seems so severe that they deem such action conceptually or psychologically impossible (Hare 1963, chap. 5; Pugmire 1982; Watson 1977).⁶ Many others try to

accommodate strict akratic action in a unified perspective (Davidson 1980, chap. 2, 1982; Dunn 1987; Mele 1987, 1995; Pears 1984).

2. EXPLAINING STRICT AKRATIC ACTION

To the extent that one's decisive judgment derives from one's motivational attitudes, it has a motivational dimension.⁷ That helps explain why many regard akratic action as theoretically perplexing. How, they wonder, can the motivation associated with a judgment of this kind be outweighed by competing motivation, especially when the competing motivational attitudes—or *desires*, broadly construed—have been taken into account in arriving at the judgment?

Elsewhere (Mele 1987) I defended an answer to this question that rests partly on two theses, both of which I defended.

1. Decisive judgments normally are formed at least partly on the basis of our evaluation of the “objects” of our desires (i.e., the desired items).
2. The motivational force of our desires does not always match our evaluation of the objects of our desires. (Santas 1966, Smith 1992, Stocker 1979, Watson 1977)⁸

If both theses are true, it should be unsurprising that sometimes, although we decisively judge it better to *A* than to *B*, we are more strongly motivated to *B* than to *A*. Given how our motivation stacks up, it should also be unsurprising that we *B* rather than *A*.

Thesis 1 is a major plank in a standard conception of practical reasoning. In general, when we reason about what to do, we inquire about what it would be best, or better, or “good enough” to do, not about what we are most strongly motivated to do. When we ask such questions while having conflicting desires, our answers typically rest significantly on our assessments of the objects of our desires—which may be out of line with the motivational force of those desires, if thesis 2 is true.

Thesis 2, as I argued in Mele 1987, is confirmed by common experience and thought experiments and has a foundation in empirical studies. Desire-strength is influenced not only by our evaluation of the objects of desires, but also by such factors as the perceived proximity of prospects for desire-satisfaction, the salience of desired objects in perception or in imagination, and the way we attend to desired objects (Ainslie 1992, Metcalfe and Mischel 1999, Rorty 1980a). Factors such as these need not have a matching effect on assessment of desired objects.

A few hours ago, an agent decisively judged it better to *A* than to *B*, but he

now has a stronger desire to *B* than to *A*. Two versions of the case merit attention. In one, along with the change in desire strength, there is a change of judgment. For example, last night, after much soul-searching, Al formed a decisive judgment favoring not eating after-dinner snacks for the rest of the month and desired more strongly to forego them than to indulge himself; but now, a few hours after dinner, Al's desire for a snack is stronger than his desire for the rewards associated with not snacking, and he decisively judges it better to have a snack than to refrain. In another version of the case, the change in relative desire strength is not accompanied by a change of judgment. Al retains the decisive judgment favoring not eating after dinner, but he eats anyway. Assuming that Al eats intentionally and is not compelled to eat, this is a strict akratic action.

Empirical studies of the role of representations of desired objects in impulsive behavior and delay of gratification (reviewed in Mele 1987, 88–93; see Mischel et al. 1989 for an overview) provide ample evidence that our representations of desired objects have two important dimensions, a motivational and an informational one. Our decisive judgments may be more sensitive to the informational dimension of our representations than to the motivational dimension, with the result that such judgments sometimes recommend courses of action that are out of line with what we are most strongly motivated to do at the time. If so, strict akratic action is a real possibility—provided that at least some intentional actions that conflict with agents' decisive judgments at the time of action are not *compelled*.

A discussion of compulsion would lead quickly to the issue of free will, which is well beyond the scope of this chapter. It is worth noting, however, that unless a desire is irresistible, it is up to the agent, in some sense, whether she acts on it. This idea is an element of both the motivational and the intellectual perspective on intentional action. Another element is the idea that relatively few desires are irresistible. Of course, a proper appreciation of the latter idea would require an analysis of irresistible desire.⁹ It may suffice for present purposes to suggest that, often, when we act against our decisive judgments, we could have used our resources for self-control in effectively resisting temptation.¹⁰ Normal agents can influence the strength of their desires in a wide variety of ways (Ainslie 1992, Metcalfe and Mischel 1999, Mischel et al. 1989). For example, they can refuse to focus their attention on the attractive aspects of a tempting course of action and concentrate instead on what is to be accomplished by acting as they judge best. They can attempt to augment their motivation for performing the action judged best by promising themselves rewards for doing so. They can picture a desired item as something unattractive—for example, a chocolate pie as a plate of chocolate-coated chewing tobacco—or as something that simply is not arousing. Desires typically do not have immutable strengths, and the plasticity of motivational strength is presupposed by standard conceptions of self-control. Occasionally we *do not* act as we judge best, but it is implausible that, in all such cases, we *cannot*

act in accordance with these judgments. (This suggestion is defended in Mele 1987, chap. 2; 1995, chap. 3; and 2002.)¹¹

3. KINDS OF AKRATIC ACTION AND THE IRRATIONALITY OF STRICT AKRATIC ACTION

Not all akratic action is of the strict kind. In this section, without aspiring to be exhaustive, I identify two additional kinds discussed in the literature. I also comment on the question of whether strict akratic action is necessarily irrational.

Socrates, in defending the thesis that no one *knowingly* does wrong, argues that what actually happens in apparent instances of strict akratic action is that, owing to the proximity of anticipated pleasures, agents change their minds about what it would be best to do (Plato 1953 355d–357d). Even if he mistakenly denies the reality of strict akratic action, Socrates identifies an important phenomenon. Some such changes of mind are *motivationally biased* processes (on motivated bias, see secs. 4–6). When one is tempted to do something that conflicts with one’s decisive judgment, one has motivation to believe that the tempting option would be best. After all, acquiring that belief would diminish one’s resistance to acting as one is tempted to act (Pears 1984, 12–13). In what I elsewhere (Mele 1996) called “Socratic akratic action” (without meaning to suggest that Socrates regarded the episodes as akratic), the agent’s new belief issues from a process biased by a desire for the tempting option. In the context of *akrasia*, the most relevant standards for determining bias are the agent’s. Perhaps the agent accepts a principle about beliefs that is violated by his present change of mind—for example, the principle that it is best not to allow what one wants to be the case to shape what one believes is the case. And if a motivationally biased change of mind of this kind is avoidable by the agent by means of an exercise of self-control, it is itself an *akratic* episode, an episode manifesting *akrasia* or an associated imperfection, for no perfectly self-controlled person makes motivated judgments that are biased relative to his own standards, if he can avoid doing so by exercising self-control. Furthermore, an intentional action that accords with the new judgment is derivatively akratic (Mele 1987, 6–7, 1996; Pears 1984, 12–13; Rorty 1980b).

Seemingly, there also are “unorthodox” instances of akratic action, in which agents act *in accordance with* their decisive judgments, and, similarly, unorthodox exercises of self-control in support of conduct that conflicts with the agents’ decisive judgments (Bigelow, Dodds, and Pargetter 1990, 46; Hill 1986, 112; Jackson

1984, 14; Kennett 2000, 120–24; Mele 1987, 7–8, 1995, 60–76). Here is an illustration of unorthodox akratic action from Mele 1995: “Young Bruce has decided to join some wayward Cub Scouts in breaking into a neighbor’s house, even though he decisively judges it best not to do so. . . . At the last minute, Bruce refuses to enter the house and leaves the scene of the crime. His doing so because his decisive judgment has prevailed is one thing; his refusing to break in owing simply to a failure of nerve is another. In the latter event, Bruce arguably has exhibited weakness of will: he ‘chickened out’ ” (60). If, instead, Bruce had mastered his fear and participated in the crime, we would have an unorthodox exercise of self-control. John Bigelow, Susan Dodds, and Robert Pargetter (1990) regard unorthodox episodes of these kinds as support for the idea that what is essential to akratic action is the presence of a second-order desire that either loses or wins against a first-order desire in the determination of action, independently of what (if anything) the agent judges it best to do.¹² For argumentation to the contrary, see Mele 1995, chap. 4.

I suggested that an agent who acts akratically against a rationally formed decisive judgment acts in a subjectively irrational way. But suppose his judgment is irrational or formed on the basis of reflection that does not take into account important, relevant attitudes of his. And suppose he acts on the basis of attitudes that constitute or reflect better reasons than the ones that ground his judgment. In that case, some have argued, his akratic action is rational, or less irrational than an action in accordance with his decisive judgment would have been (Arpaly 2000, Audi 1990, McIntyre 1990). In such cases, an akratic action may reflect an agent’s system of values better than his decisive judgment does.¹³ However, the fact that a certain akratic action against a particular rationally formed decisive judgment is more coherent with the agent’s system of values or reasons than an action in accordance with that judgment would be is consistent with the akratic action’s being subjectively irrational, to some degree, in virtue of failing to cohere with the judgment. (The same may be said of akratic actions against decisive judgments that are irrationally formed.) Also, it may be doubted that the problem with a rationally formed decisive judgment that is not sensitive to certain of the agent’s values or reasons is irrationality.

4. MOTIVATED IRRATIONAL BELIEF: AGENCY AND ANTI-AGENCY VIEWS

“Biased” or “irrational” beliefs are biased or irrational relative to some standard or other. In the context of akratic belief (Davidson 1985; Heil 1984; Mele 1987,

chap. 8; Pears 1984, chap. 4; Rorty 1983), the germane standards are the believer's own. In work on self-deception, general epistemic standards are typically assumed. One test for motivationally biased belief of a sort appropriate to self-deception is the following. If S is self-deceived in believing that 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) and they were to engage in at least as much reflection on the issue as S does and at least a moderate amount of reflection, those who conclude that p is false would significantly outnumber those who conclude that p is true (cf. Mele 2001, 106). Two plausible requirements for impartiality in this context are that one neither desire that p nor desire that $\sim p$ and that one not prefer avoidance of either of the following errors over the other: falsely believing that p and falsely believing that $\sim p$.

The question whether all motivationally biased beliefs are irrational raises an intriguing question about rationality. Might a motivationally biased belief be rational—or, at least, not irrational—from some legitimate point of view? W. K. Clifford asserts that “it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence” (1886, 346). William James denies this, contending that “a rule of thinking which would absolutely prevent me from acknowledging certain kinds of truth if those kinds of truth were really there, would be an irrational rule” ([1897] 1979, 31–32). It is at least consistent with James's claim that a person who is rightly convinced that he would be miserable if he were no longer to believe that God exists may rationally believe, on insufficient evidence, that God exists. I set this question about rationality aside and pursue the issue of motivationally biased beliefs. These beliefs are irrational by general epistemic standards or from an epistemic point of view.

Consider two bold theses about motivationally biased beliefs:

1. *The agency view.* 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 try to make it easier for ourselves to acquire or retain it.
2. *The anti-agency view.* 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 it.

Probably, the truth lies somewhere between these poles. But which thesis is closer to the truth?

One problem for the agency view is central to a familiar puzzle about self-deception. The attempts to which the view appeals threaten to undermine themselves. If Al is trying to bring it about that he believes that he is a good day trader—not by improving his investment skills, but by ignoring or downplaying evidence that he is an inferior trader while searching for evidence that he has superior investment skills—won't he see that the “grounds” for belief that he arrives at in this way are illegitimate? And won't he consequently fail in his at-

tempt? A predictable reply is that the “tryings” or efforts at issue are not conscious efforts and therefore need not block their own success in this way.¹⁴ 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 anti-agency view also is linked to this 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 Al’s wanting it to be the case that he is a good day trader motivate him to believe that he is good at this except by motivating him to try to bring it about that he believes this or to try to make it easier for himself to believe this?¹⁵ The anti-agency view is faced with a clear challenge: to provide an alternative account of the mechanism(s) by which desires lead to motivationally biased beliefs.

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, 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 1985b, 145–46)

Both views rest mainly on the thought that the only, or best, way to account for certain data is to hold that the person, or some center of agency within her, tries to bring it about that she, or some “system” in her, holds a certain belief.

Consider a case of self-deception similar to the one Davidson diagnoses in the passage last quoted. Carlos “has good reason to believe” that he will fail his driver’s test (Davidson 1985b, 145). “He has failed . . . twice . . . and his instructor has said discouraging things. On the other hand, he knows the examiner personally, and he has faith in his own charm. . . . The thought of failing . . . again is painful” (145–46). Suppose the overwhelming majority of Carlos’s impartial cognitive peers presented with his evidence would believe that Carlos will fail and none would believe that he will pass. (Some peers with high standards for belief may withhold belief.) Even so, in the face of the contrary evidence, Carlos believes that he will pass. Predictably, he fails.

Self-deception is often thought to be such that if Carlos is self-deceived in believing that he will pass the test, he believed at some time that he would fail it (Bach 1981, Demos 1960, Haight 1980, Quattrone and Tversky 1984, Rorty 1972, Sackeim and Gur 1985). However, in accommodating the data offered about Carlos, there is no evident need to suppose he had this true belief. Perhaps his self-deception is such that not only does he acquire the belief that he will pass, but he never acquires the belief that he will fail. In fact, this seems true of much self-deception. Seemingly, 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 never 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, there probably 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 some such exercise of agency 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 so straightforward.

It is often held that, in Jeffrey Foss's words, "desires have no explanatory force without associated beliefs" that identify (apparent) means to the desires' satisfaction, and this is part of "the . . . logic of belief-desire explanation" (1997, 112). This claim fares poorly in the case of motivationally biased belief. "A survey of one million high school seniors found that . . . 25% thought they were in the top 1%" in ability to get along with others (Gilovich 1991, 77). A likely hypothesis about this striking figure includes the idea that desires that p can contribute to biased beliefs that p . If Foss's claim were true, a student's wanting it to be true that she is exceptionally affable would help explain her believing that she is only in combination with an instrumental belief (or a collection thereof) that links her believing that she is superior in this sphere to the satisfaction of her desire to be superior. But we search in vain for instrumental beliefs that are plausibly regarded as turning the trick frequently enough to accommodate the data. Perhaps believing that one is exceptionally affable can help bring it about that one is superior in this sphere, and some high school students may believe that this is so. But it is highly unlikely that most people who have a motivationally biased belief that they are exceptionally affable have that 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. No other instrumental beliefs look more promising.

Should we infer, then, that wanting it to be true that one has a superior

ability to get along with others plays a role in explaining only relatively few unwarranted beliefs that one is superior in this area? Not at all. There is powerful empirical evidence, some of which is reviewed shortly, 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 they can have.

5. EVIDENCE FOR AND SOURCES OF MOTIVATIONALLY BIASED BELIEF

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%." 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 Foss's claim about the explanatory force of desires will be inclined toward some version or other of the agency view of motivationally biased belief and self-deception. However, as I will explain, 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.

The survey I mentioned also found that 70 percent of high school seniors "thought they were above average in leadership ability, and only 2% thought they were below average." "A survey of university professors found that 94% thought they were better at their jobs than their average colleague" (Gilovich 1991, 77). Data such as these suggest that desires sometimes bias beliefs. The aggregated self-assessments are wildly 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. There is powerful evidence that people have a tendency to believe propositions that they want to be true even when an impartial investigation of readily available data would indicate that they probably are false. A plausible hypothesis about that tendency is that desire sometimes biases belief.

Controlled studies provide confirmation for this hypothesis. In one study, 75 women and 86 men read an article asserting 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 can-

cer”; and that “caffeine induced the disease by increasing the concentration of a substance called cAMP in the breast” (Kunda 1987, 642). (Because the article did not directly 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” (643–44). Female “heavy consumers” of caffeine were significantly less convinced of the connections than female “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 the female “heavy consumers” were the most seriously threatened by it, a plausible hypothesis is that a desire that their coffee drinking has not significantly endangered their health helps to account for their lower level of conviction (Kunda 1987, 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.” Along with the lesser threat, there is less motivation for skepticism about the evidence.

Attention to some phenomena that have been argued to be sources of *unmotivated* biased belief sheds light on motivationally biased belief. A number of such sources have been identified, including the following two.

1. *Vividness of information.* A datum’s vividness for us often is a function of such things as its concreteness, its “imagery-provoking” power, and its sensory, temporal, or spatial proximity (Nisbett and Ross 1980, 45). Vivid data are more likely to be recognized, attended to, and recalled than pallid data. Consequently, vivid data tend to have a disproportional influence on the formation and retention of beliefs.
2. *The confirmation bias.* People testing a hypothesis tend to search (in memory and the world) more often for confirming than for disconfirming instances and to recognize the former more readily (Baron 1988, 259–65; Klayman and Ha 1987; Nisbett and Ross 1980, 181–82). This is true even when the hypothesis is only a tentative one (as opposed, e.g., to a belief one has). People also tend to interpret relatively neutral data as supporting a hypothesis they are testing (Trope, Gervy, and Liberman 1997, 115).

Although sources of biased belief apparently can function independently of motivation, they also may be triggered and sustained by desires in the production of *motivationally* biased beliefs.¹⁶ For example, desires can enhance the vividness or salience of data. Data that count in favor of the truth of a proposition that one hopes is true may be rendered more vivid or salient by one’s recognition that they so count. Similarly, desires can influence which hypotheses occur to one and

affect the salience of available hypotheses, thereby setting the stage for the confirmation bias.¹⁷ Owing to a desire that p , one may test the hypothesis that p is true rather than the contrary hypothesis. In these ways and others, a desire that p may help explain the acquisition of an unwarranted belief that p .

Sometimes we generate our own hypotheses, and sometimes others suggest hypotheses to us—including extremely unpleasant ones. If we were always to concentrate primarily on confirmation in hypothesis testing, independently of what is at stake, that would indicate the presence of a cognitive tendency or disposition that uniformly operates independently of desires and that desires never play a role in influencing the proportion of attention we give to evidence for the falsity of a hypothesis. However, there is powerful evidence that the “confirmation bias” is much less rigid than this. For example, in one study (Gigerenzer and Hug 1992), two groups of subjects were asked to test “social-contract rules such as If someone stays overnight in the cabin, then that person must bring along a bundle of firewood . . . ” (Friedrich 1993, 313). The group asked to adopt “the perspective of a cabin guard monitoring compliance” showed an “extremely high frequency” of testing for disconfirmation (i.e., for visitors who stay in the cabin overnight but bring no wood). The other group, asked to “take the perspective of a visitor trying to determine” whether firewood was supplied by visitors or by a local club, displayed the common confirmation bias.¹⁸

6. A MOTIVATIONAL MODEL OF LAY HYPOTHESIS TESTING

An interesting recent theory of lay hypothesis testing is designed, in part, to accommodate data of the sort I have been describing. I explored it in Mele 2001, where I offered grounds for caution and moderation and argued that a qualified version is plausible.¹⁹ I named it the “FTL theory,” after the authors of the two essays on which I primarily drew, Friedrich 1993 and Trope and Liberman 1996. Here, I offer a thumbnail sketch.

The basic idea of the FTL theory is that a concern to minimize costly errors drives lay hypothesis testing. The *errors* on which the theory focuses are false beliefs. The *cost* of a false belief is the cost, including missed opportunities for gains, that it would be reasonable for the person to expect the belief—if false—to have, given his desires and beliefs, if he were to have expectations about such things. A central element of the FTL theory is a “confidence threshold”—or a “threshold,” for short. The lower the threshold, the thinner the evidence sufficient

for reaching it. Two thresholds are relevant to each hypothesis: “The acceptance threshold is the minimum confidence in the truth of a hypothesis,” p , sufficient for acquiring a belief that p “rather than continuing to test [the hypothesis], and the rejection threshold is the minimum confidence in the untruth of a hypothesis,” p , sufficient for acquiring a belief that p “and discontinuing the test” (Trope and Liberman 1996, 253). The two thresholds often are not equally demanding, and acceptance and rejection thresholds respectively depend “primarily” on “the cost of false acceptance relative to the cost of information” and “the cost of false rejection relative to the cost of information.” The “cost of information” is simply the “resources and effort” required for gathering and processing “hypothesis-relevant information” (252).

Confidence thresholds are determined by the strength of aversions to specific costly errors together with information costs. Setting aside the latter, the stronger one’s aversion to falsely believing that p , the higher one’s threshold for belief that p . These aversions influence belief in a pair of related ways. First, because, other things being equal, lower thresholds are easier to reach than higher ones, belief that $\sim p$ is a more likely outcome than belief that p , other things being equal, in a hypothesis tester who has a higher acceptance threshold for p than for $\sim p$. Second, the aversions influence *how* we test hypotheses, not just *when we stop* testing them (owing to our having reached a relevant threshold). Recall the study in which subjects asked to adopt “the perspective of a cabin guard” showed an “extremely high frequency” of testing for disconfirmation, whereas subjects asked to “take the perspective of a visitor” showed the common confirmation bias.

It might be claimed that if aversions to specific errors function in the second way just identified, they work together with beliefs to the effect that testing-behavior of a particular kind is conducive to avoiding these errors. It might be claimed, accordingly, that the pertinent testing-behavior is performed with the intention of avoiding, or of trying to avoid, the pertinent error. The thrust of these claims is that the FTL theory accommodates the confirmation bias, for example, by invoking a model of intentional action.

This is not a feature of the FTL model, as its proponents understand it. Friedrich, for example, claims that desires to avoid specific errors can trigger and sustain “automatic test strategies” (313), which supposedly happens in roughly the nonintentional way in which a desire that p results in the enhanced vividness of evidence for p . In Mele 2001 (41–49, 61–67), I argued that a person’s being more strongly averse to falsely believing that $\sim p$ than to falsely believing that p may have the effect that he primarily seeks evidence for p , is more attentive to such evidence than to evidence that $\sim p$, and interprets relatively neutral data as supporting p , without this effect’s being mediated by a belief that such behavior is conducive to avoiding the former error. The stronger aversion may simply frame the topic in such a way as to trigger and sustain these manifestations of the confirmation bias without the assistance of a belief that behavior of this kind is

a means of avoiding particular errors. Similarly, having a stronger aversion that runs in the opposite direction may result in a skeptical approach to hypothesis testing that in no way depends on a belief to the effect that an approach of this kind will increase the probability of avoiding the costlier error. Given the aversion, skeptical testing is predictable independently of the agent's believing that a particular testing style will decrease the probability of making a certain error.

The FTL theory applies straightforwardly to both "straight" and "twisted" self-deception (Mele 2001, 4–5, 94–118). In straight cases, we are self-deceived in believing something that we want to be true. In twisted cases, we are self-deceived in believing something that we want to be *false* (and do not also want to be true). Twisted self-deception may be exemplified by an insecure, jealous husband who believes that his wife is having an affair despite possessing only relatively flimsy evidence for that proposition and despite unambivalently wanting it to be false that she is so engaged.²⁰ Friedrich writes: "A prime candidate for primary error of concern is believing as true something that leads [one] to mistakenly criticize [oneself] or lower [one's] self-esteem. Such costs are generally highly salient and are paid for immediately in terms of psychological discomfort. When there are few costs associated with errors of self-deception (incorrectly preserving or enhancing one's self-image), mistakenly revising one's self-image downward or failing to boost it appropriately should be the focal error" (314). Here, he plainly has straight self-deception in mind.

Whereas, for many people, it may be more important to avoid acquiring the false belief that their spouses are having affairs than to avoid acquiring the false belief that they are not so engaged, the converse may well be true of some insecure, jealous people. The belief that one's spouse is unfaithful tends to cause significant psychological discomfort. Even so, avoiding falsely believing that their spouses are faithful may be so important to some people that they test relevant hypotheses in ways that, other things being equal, are less likely to lead to a false belief in their spouses' fidelity than to a false belief in their spouses' infidelity. Furthermore, data suggestive of infidelity may be especially salient for these people and contrary data quite pallid by comparison. Don Sharpsteen and Lee Kirkpatrick observe that "the jealousy complex"—that is, "the thoughts, feelings, and behavior typically associated with jealousy episodes"—is interpretable as a mechanism "for maintaining close relationships" and appears to be "triggered by separation, or the threat of separation, from attachment figures" (1997, 627). It certainly is conceivable that, given a certain psychological profile, a strong desire to maintain one's relationship with one's spouse plays a role in rendering the potential error of falsely believing one's spouse to be innocent of infidelity a "costly" error, in the FTL sense, and more costly than the error of falsely believing one's spouse to be guilty. After all, the former error may reduce the probability that one takes steps to protect the relationship against an intruder. The FTL theory

provides a basis for a plausible account of twisted self-deception (Mele 2001, chap. 5).

7. CONCLUSION: THE PARADOX OF IRRATIONALITY

Donald Davidson writes: “The underlying paradox of irrationality, from which no theory can entirely escape, is this: if we explain it too well, we turn it into a concealed form of rationality; while if we assign incoherence too glibly, we merely compromise our ability to diagnose irrationality by withdrawing the background of rationality needed to justify any diagnosis at all” (1982, 303). The explanations sketched here of strict akratic action and motivationally biased belief avoid Davidson’s worries about paradox. Akratic agents act for reasons, and in central cases, they make rational decisive judgments: “the background of rationality” required for that is in place. But insofar as their uncompelled actions are at odds with their rational decisive judgments, they act irrationally. Motivationally biased believers test hypotheses and believe on the basis of evidence. Again there is a background of rationality. But, owing to the influence of motivation, they violate general standards of epistemic rationality.

NOTES

Parts of this chapter derive from Mele 1987, 1995, 1998, and 2001. I am grateful to Piers Rawling for comments on a draft.

1. For a detailed account, see Mele 1995, chaps. 1–7.
2. Assuming a middle ground between *akrasia* and self-control, not all akratic actions manifest *akrasia*. Someone who is more self-controlled than most people in a certain sphere may, in a particularly trying situation, succumb to temptation in that sphere against her better judgment. If her intentional action is uncompelled, she has acted akratically—even if her action manifests not *akrasia* but an associated imperfection.
3. An agent who makes such a judgment may or may not proceed to search for additional options. He may regard the best member of his currently envisioned options as “good enough.”
4. On failures of coherence, see Arpaly 2000 and Harman, chap. 3, this volume.
5. On the nature of motivational strength and the theoretical utility of the notion, see Mele 2003a, chap. 7.

6. For replies to Hare, Pugmire, and Watson, see Mele 1987, chap. 2 and 51–55. See also Pugmire 1994, responding to Mele 1987, and the rejoinder in Mele 1995, 44–54.
7. This is not to say that motivation is “built into” the judgment itself.
8. For opposition to the idea that desires vary in motivational strength, see Charlton 1988, Gosling 1990, and Thalberg 1985. For a reply to the opposition, see Mele 2003a, chap. 7.
9. See Mele 1992, chap. 5, for an analysis of irresistible desire.
10. This is not a necessary condition for strict akratic action. There are Frankfurt-style cases (Frankfurt 1969) in which, although one *A*-ed akratically and without any external interference, if one had been about to resist temptation, a mind-reading demon would have prevented one from doing so (see Mele 1995, 94–95).
11. One might claim that anyone who is more strongly motivated to *B* than to *A* will also be more strongly motivated to allow that feature of her motivational condition to persist than to change it. I rebut this claim in Mele 1987, chap. 6, and Mele 1995, chap. 3.
12. For similar positions on akratic action, see Schiffer 1976 and Swanton 1992, chap. 10. Swanton contends that “in the context of weakness of will, the will should be identified with strong evaluation,” a certain kind of “evaluative second-order desire” (149). Also see Jeffrey 1974. On an alternative view, the agent who akratically *A*-s believes that she should believe that it is best not to *A* but does not believe what she believes she should (Tenenbaum 1999; cf. Buss 1997, 36).
13. Dion Scott-Kakures (1997) argues that akratic agents are wrong about what they have “more reason” to do.
14. See, e.g., Bermudez 2000, Quattrone and Tversky 1984, Sackeim 1988, and Talbott 1995. A related response is mental partitioning: the deceived part of the mind is unaware of what the deceiving part is up to. See Pears 1984 (cf. 1991) for a detailed response of this kind and Davidson 1985 (cf. 1982) for more modest partitioning. For criticism of partitioning views of self-deception, see Barnes 1997, Johnston 1988, and Mele 1987.
15. Two uses of “motivate” should be distinguished. In one, a desire’s motivating an action or a belief is a matter of its *being* motivation for it. Piers’s desire to fish today is motivation for him to fish, even if, desiring more to work on his chapter, he foregoes a fishing trip. In another use, a desire motivates something only if, in addition to being motivation for it, it plays a role in *producing* that thing. Here, I use “motivate” in the second sense.
16. I develop this idea in Mele 1987, chap. 10, and Mele 2001. Kunda 1990 develops the same theme, concentrating on evidence that motivation sometimes primes the confirmation bias. Also see Kunda 1999, chap. 6.
17. For motivational interpretations of the confirmation bias, see Friedrich 1993 and Trope and Liberman 1996, 252–65.
18. For further discussion, see Samuels and Stich, chap. 15, this volume.
19. See Mele 2001, 31–49, 63–70, 90–91, 96–98, 112–18.
20. On this case, see Barnes 1997, chap. 3; Lazar 1999, 274–77; and Pears 1984, 42–44. Also see Davidson 1985, 144; Demos 1960, 589; McLaughlin 1988, 40; Mele 1987, 114–18; and Mele 2001, chap. 5.