BEHAVIORAL AND BRAIN SCIENCES (1997) **20,** 91–136
Printed in the United States of America

Real self-deception

Alfred R. Mele
Department of Philosophy, Davidson College, Davidson, NC 28036.
Electronic mail: *almele@davidson.edu*

Abstract: Self-deception is made unnecessarily puzzling by the assumption that it is an intrapersonal analog of ordinary interpersonal deception. In paradigmatic cases, interpersonal deception is intentional and involves some time at which the deceiver disbelieves what the deceived believes. The assumption that self-deception is intentional and that the self-deceiver believes that some proposition is true while also believing that it is false produces interesting conceptual puzzles, but it also produces a fundamentally mistaken view of the dynamics of self-deception. This target article challenges the assumption and presents an alternative view of the nature and etiology of self-deception. Drawing upon empirical studies of cognitive biases, it resolves familiar "paradoxes" about the dynamics of self-deception and the condition of being self-deceived. Conceptually sufficient conditions for self-deception are offered and putative empirical demonstrations of a kind of self-deception in which a subject believes that a proposition is true while also believing that it is false are criticized. Self-deception is neither irresolvably paradoxical nor mysterious, and it is explicable without the assistance of mental exotica. The key to understanding its dynamics is a proper appreciation of our capacity for acquiring and retaining motivationally biased beliefs.

Keywords: belief; bias; consciousness; contradictory beliefs; intention; motivation; rationality; self-deception; wishful thinking

1. Introduction

Self-deception poses tantalizing conceptual conundrums and provides fertile ground for empirical research. Recent interdisciplinary volumes on the topic feature essays by biologists, philosophers, psychiatrists, and psychologists (Lockard & Paulhus 1988; Martin 1985). Self-deception's location at the intersection of these disciplines is explained by its significance for questions of abiding interdisciplinary interest: To what extent is our mental life present – or even accessible – to consciousness? How rational are we? How is "motivated irrationality" to be explained? To what extent are our beliefs subject to our control? What are the determinants of belief? How does motivation bear upon belief? To what extent are widely shared psychological tendencies products of evolution?[1]

A proper grasp of the dynamics of self-deception may yield substantial *practical* gains. Plato wrote, "There is nothing worse than self-deception – when the deceiver is at home and always with you" (*Cratylus* 428d). Others argue that self-deception is sometimes beneficial; whether we would be better or worse off, on the whole, if we never deceived ourselves is an open question.[2] In any case, ideally, a detailed understanding of the etiology of self-deception would help reduce the frequency of harmful self-deception. This hope is boldly voiced by Jonathan Baron in a book on rational thinking and associated obstacles: "If people *know* that their thinking is poor, they will not believe its results. One of the purposes of a book like this is to make recognition of poor thinking more widespread, so that it will no longer be such a handy means of self-deception" (Baron 1988, p. 39). A lively debate in social psychology about the extent to which sources of biased belief are subject to personal control has generated evidence that some prominent sources of bias are to some degree controllable.[3] This provides grounds for hope that a better understanding of self-deception would enhance our ability to do something about it.

My aim in this target article is to clarify the nature and (relatively proximate) etiology of self-deception. Theorists have tended to construe self-deception as largely isomorphic with paradigmatic interpersonal deception. Such construals, which have generated some much-discussed puzzles or "paradoxes," guide influential work on self-deception in each of the four disciplines mentioned (e.g., Davidson 1985; Gur & Sackeim 1979; Haight 1980; Pears 1984; Quattrone & Tversky 1984; Trivers 1985).[4] In the course of resolving the major puzzles, I will argue that the attempt to understand self-deception on the model of paradigmatic interpersonal deception is fundamentally misguided. Section 2 provides background, including sketches of two familiar puzzles: one about the mental state of a self-deceived person at a given time, the other about the dynamics of self-deception. Section 3, drawing upon empirical studies of biased belief, resolves the first puzzle and articulates sufficient conditions for self-deception. Section 4 challenges some attempted empirical demonstrations of the reality of self-deception, construed as requiring the simultaneous possession of beliefs whose propositional contents are mutually contradictory. Section 5 resolves the dynamic puzzle. Section 6 examines *intentional* self-deception.

Readers should be forewarned that the position 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 fun, the main source of broader, enduring interest in self-deception is a desire to understand and explain the behavior of real human beings.

© *1997 Cambridge University Press* *0140-525X/97 $9.00+.10*

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

2. Three approaches to characterizing self-deception and a pair of puzzles

Defining "self-deception" is no mean feat. Three common approaches may be distinguished. One is *lexical:* a theorist starts with a definition of "deceive" or "deception," using the dictionary or common usage as a guide, and then uses it as a model for defining self-deception. Another is *example-based:* one scrutinizes representative examples of self-deception and attempts to identify their essential common features. The third is *theory-guided:* 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 common-sense explanatory hypotheses that may be misguided: ordinary folks may be good at identifying hypothetical cases of self-deception but quite unreliable at diagnosing what transpires 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, as I will explain. Lexicalists need a sense of the word 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 common-sense theories about what transpires 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 are 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 $\sim p$ (i.e., that *p* is false) 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 know, or at least believe truly, that $\sim p$ and cause oneself to believe that *p*. At the very least, one starts out believing that $\sim p$ 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 $\sim p$ (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. 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.*

Assumption 2 generates a *dynamic* puzzle, a puzzle about the dynamics of self-deception. It is often held that doing something intentionally entails doing it knowingly. If that is so, and if *deceiving* is by definition an intentional activity, then one who deceives oneself does so *knowingly.* But knowingly deceiving oneself into believing that *p* would require knowing that what one is getting oneself to believe is false. How can that knowledge fail to undermine the very project of deceiving oneself? 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.[5] Furthermore, deception is normally 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?

In sketching these puzzles, I conjoined the numbered assumptions with subsidiary ones. One way for a proponent of the reality of self-deception to attempt to solve the puzzles is to attack the subsidiary assumptions while leaving the main assumptions unchallenged. A more daring tack is to undermine the main assumptions, 1 and 2. That is the line I will 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 $\sim p$. 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" (my authority is 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 $\sim p$. Yesterday, mistakenly believing that my son's keys were on my desk, I told him they were there. In so doing, I caused him to believe a falsehood. I deceived him, in the sense identified; but I did not do so intentionally, nor did I cause him 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 transpires in such cases that these people start by knowing or believing the truth, *p,* and intentionally cause themselves to believe that $\sim p$? 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.

Compare the question whether self-deception is properly understood on the model of interpersonal deception with the question whether addiction is properly understood on the model of disease. Perhaps the current folk-conception of addiction treats addictions as being

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

diseases by definition. However, the disease model of addiction has been forcefully attacked (e.g., Peele 1989). The issue is essentially about explanation, not semantics. How is the characteristic behavior of people typically counted as addicts best explained? Is the disease model of addiction more fruitful for explanation than its competitors? Self-deception, like addiction, is an explanatory concept. We postulate self-deception in particular cases to explain behavioral data. And we should ask how self-deception is likely to be constituted – what it is likely to be – if it helps to explain the relevant data. Should we discover that the behavioral 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.

A caveat is in order before I move on. In the literature on self-deception, "belief" rather than "degree of belief" is usually the operative notion. Here, I follow suit, primarily 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)."

3. Motivated belief and the static puzzle

In stock examples of self-deception, people typically believe something they *want* to be true: that their spouses are not involved in extramarital flings, that their children are not using drugs, and so on. It is a commonplace that self-deception, in garden-variety cases, is *motivated* by wants such as these.[6] Should it turn out that the motivated nature of self-deception entails that self-deceivers intentionally deceive themselves and requires that those who deceive themselves into believing that *p* start by believing that $\sim p$, theorists who seek a tight fit between self-deception and stereotypical interpersonal deception would be vindicated. Whether self-deception can be motivated without being intentional – and without the self-deceiver's starting with the relevant true belief – remains to be seen.

A host of studies have produced results that are utterly unsurprising on the hypothesis that motivation sometimes biases beliefs. Thomas Gilovich (1991) 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. (p. 77)

Apparently, we have a tendency to believe propositions 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 *wanting* something to be true sometimes exerts a biasing influence on what we believe.

Ziva Kunda, in a recent review essay, 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 are constructed so as to justify desired conclusions" (Kunda 1990, p. 483). In an especially persuasive study, undergraduate subjects (75 women and 86 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). (Since 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).

Given that all subjects were exposed to the same information and assuming that only the female "heavy consumers" were personally threatened by it, a plausible hypothesis is that their lower level of conviction is due to "motivational processes designed to preserve optimism about their future health" (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). 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.[7] In either case, however, one cannot 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 get themselves to find it less convincing. Motivation can prompt cognitive behavior protective of favored beliefs without the person's intending to protect those beliefs. Many instances of self-deception, as I will argue, are explicable along similar lines.

Beliefs that we are self-deceived in acquiring or retaining are a species of *biased* belief. In self-deception, on a widely held view, the biasing is *motivated*. Even so, attention to some sources of *unmotivated* or "cold" biased belief will prove salutary. A number of such sources have been identified in psychological literature. Here are four.[8]

3.1.1. Vividness of information. A datum's vividness for an individual is often a function of individual interests, the concreteness of the datum, its "imagery-provoking" power, or its sensory, temporal, or spatial proximity (Nisbett &

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Ross 1980, p. 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.[9]

3.1.2. The availability heuristic. When we form beliefs about the frequency, likelihood, or causes of an event, we "often may be influenced by the relative availability of the objects or events, that is, their accessibility in the processes of perception, memory, or construction from imagination" (Nisbett & Ross 1980, p. 18). For example, we may mistakenly believe that the number of English words beginning with "r" greatly outstrips the number having "r" in the third position, because we find it much easier to produce words on the basis of a search for their first letter (Tversky & Kahnemann 1973). Similarly, attempts to locate the cause(s) of an event are significantly influenced by manipulations that focus one's attention on a potential cause (Nisbett & Ross 1980, p. 22; Taylor & Fiske 1975; 1978).

3.1.3. 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, pp. 259–65; Nisbett & Ross 1980, pp. 181–82). This is true even when the hypothesis is only a tentative one (as opposed, e.g., to a belief one has). The implications of this tendency for the retention and formation of beliefs are obvious.

3.1.4. Tendency to search for causal explanations. We tend to search for causal explanations of events (Nisbett & Ross 1980, pp. 183–86). In a plausible view of the macroscopic world, this is as it should be. But given 3.1.1 and 3.1.2 above, the causal explanations upon which we so easily hit in ordinary life may often be ill-founded, and given 3.1.3, one is likely to endorse and retain one's first hypothesis much more often than one ought. Furthermore, ill-founded causal explanations can influence future inferences.

Obviously, the most vivid or available data sometimes have the greatest evidential value; the influence of such data is not *always* a biasing influence. The main point to be made is that although sources of biased belief can function independently of motivation, they may also be primed by motivation in the production of particular *motivationally* biased beliefs.[10] For example, motivation can enhance the vividness or salience of certain data. Data that count in favor of the truth of a hypothesis that one would like to be true might be rendered more vivid or salient given one's recognition that they so count, and vivid or salient data, given that they are more likely to be recalled, tend to be more "available" than pallid counterparts. Similarly, motivation can influence which hypotheses occur to one (including causal hypotheses) and affect the salience of available hypotheses, thereby setting the stage for the confirmation bias.[11] When this happens, motivation brings about cognitive behavior that epistemologists shun. False beliefs produced or sustained by such motivated cognitive behavior in the face of weightier evidence to the contrary are, I will argue, beliefs that one is self-deceived in holding. And the self-deception in no way requires that the agents intend to deceive themselves, or intend to produce or sustain certain beliefs in themselves, or start by believing something they end up disbelieving. Cold biasing is not intentional, and mechanisms of the sort described may be primed by motivation independently of any intention to deceive.

There are a variety of ways in which our desiring that *p* can contribute to our believing that *p* in instances of self-deception. Here are some examples[12]:

3.2.1. Negative misinterpretation. Our desiring that *p* may lead us to misinterpret as not counting (or not counting strongly) against *p* data that we would easily recognize to count (or count strongly) against *p* in the desire's absence. For example, Don just received a rejection notice on a journal submission. He hopes that his article was *wrongly* rejected, and he reads through the comments offered. Don decides that the referees misunderstood a certain crucial but complex point and that their objections consequently do not justify the rejection. However, as it turns out, the referees' criticisms were entirely justified; and when, a few weeks later, Don rereads his paper and the comments in a more impartial frame of mind, it is clear to him that the rejection was warranted.

3.2.2. Positive misinterpretation. Our desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence. For example, Sid is very fond of Roz, a college classmate with whom he often studies. Wanting it to be true that Roz loves him, he may interpret her refusing to date him and her reminding him that she has a steady boyfriend as an effort on her part to "play hard to get" in order to encourage Sid to continue to pursue her and prove that his love for her approximates hers for him. As Sid interprets Roz's behavior, not only does it fail to count against the hypothesis that she loves him, it is evidence *for* the truth of that hypothesis.

3.2.3. Selective focusing/attending. Our desiring that *p* may lead us both to fail to focus attention on evidence that counts against *p* and to focus instead on evidence suggestive of *p*. Attentional behavior may be either intentional or unintentional. Ann may tell herself that it is a waste of time to consider her evidence that her husband is having an affair, since he loves her too much to do such a thing; and she may intentionally act accordingly. Or, because of the unpleasantness of such thoughts, Ann may find her attention shifting whenever the issue suggests itself.

3.2.4. Selective evidence-gathering. Our desiring that *p* may lead us both to overlook easily obtained evidence for $\sim p$ and to find evidence for *p* that is much less accessible. A historian of philosophy who holds a certain philosophical position hopes that her favorite philosopher (Plato) did so too; consequently, she scours the texts for evidence of this while consulting commentaries that she thinks will provide support for the favored interpretation. Our historian may easily miss rather obvious evidence to the contrary, even though she succeeds in finding obscure evidence for her favored interpretation. Selective evidence-gathering may be analyzed as a combination of "hypersensitivity" to evidence (and sources of evidence) for the desired state of affairs and "blindness" – of which there are, of course, degrees – to contrary evidence (and sources thereof).[13]

In none of the examples offered does one hold the true belief that $\sim p$ and then intentionally bring it about that one believes that *p*. Yet, assuming that my hypothetical agents acquire relevant false beliefs in the ways described, these are garden-variety instances of self-deception. Don is self-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

deceived in believing that his article was wrongly rejected, Sid is self-deceived in believing certain things about Roz, and so on.

It is sometimes claimed that while we are deceiving ourselves into believing that *p* we must be aware that our evidence favors ~*p*, on the grounds that this awareness is part of what explains our motivationally biased treatment of data (Davidson 1985, p. 146). The thought is that without this awareness we would have no reason to treat data in a biased way, since the data would not be viewed as threatening, and consequently we would not engage in motivationally biased cognition. In this view, self-deception is understood on the model of *intentional action:* the agent has a goal, sees how to promote it, and seeks to promote it in that way. However, the model places excessive demands on self-deceivers.[14] Cold or unmotivated biased cognition is not explained on the model of intentional action, and motivation can prime mechanisms for the cold biasing of data in us without our being aware, or believing, that our evidence favors a certain proposition. Desire-influenced biasing may result both in our not being aware that our evidence favors ~*p* over *p* and in our acquiring the belief that *p*. This is a natural interpretation of the illustrations I offered of misinterpretation and of selective focusing/attending. In each case, the person's evidence may favor the undesirable proposition, but there is no need to suppose the person is aware of this in order to explain the person's biased cognition.[15] Evidence that one's spouse is having an affair (or that a scholarly paper one painstakingly produced is seriously flawed, or that someone one loves lacks reciprocal feelings) may be threatening even if one lacks the belief, or the awareness, that that evidence is stronger than one's contrary evidence.

Analyzing self-deception is a difficult task; providing a plausible set of sufficient conditions for self-deception is less demanding. Not all cases of self-deception need involve the acquisition of a new belief. Sometimes we may be self-deceived in retaining a belief that we were not self-deceived in acquiring. Still, the primary focus in the literature has been on self-deceptive belief-acquisition, and I will follow suit.

I suggest that the following conditions are jointly sufficient for *entering self-deception in acquiring a belief that* p.

1. The belief that *p* which *S* acquires is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
3. This biased treatment is a nondeviant cause of *S*'s acquiring the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for ~*p* than for *p*.[16]

Each condition requires brief attention. Condition 1 captures a purely lexical point. A person is, by definition, *deceived in* believing that *p* only if *p* is *false;* the same is true of being *self-deceived in* believing that *p*. The condition in no way implies that the falsity of *p* has special importance for the *dynamics* of self-deception. Motivationally biased treatment of data may sometimes result in someone's believing an improbable proposition, *p*, that, as it happens, is *true*. There may be self-deception in such a case; but the person is not self-deceived in believing that *p*, nor in acquiring the belief that *p*.[17]

My brief discussion of various ways of entering self-deception serves well enough as an introduction to condition 2. My list of motivationally biased routes to self-deception is not intended as exhaustive, but my discussion of these routes does provide a gloss on the notion of motivationally biased treatment of data.

My inclusion of the term "nondeviant" in condition 3 is motivated by a familiar problem for causal characterizations of phenomena in any sphere (see, e.g., Mele 1992a, Ch. 11). Specifying the precise nature of nondeviant causation of a belief by motivationally biased treatment of data is a difficult technical task better reserved for another occasion. However, much of this article provides guidance on the issue.

The thrust of condition 4 is that self-deceivers believe against the weight of the evidence they possess. For reasons offered elsewhere, I do not view 4 as a *necessary* condition of self-deception (Mele 1987a, pp. 134–35). In some instances of motivationally biased evidence-gathering, for example, people may bring it about that they believe a falsehood, *p*, when ~*p* is much better supported by evidence readily available to them, even though, owing to the selectivity of the evidence-gathering process, the evidence that they themselves actually *possess* at the time favors *p* over ~*p*. As I see it, such people are naturally deemed self-deceived, other things being equal. Other writers on the topic do require that a condition like 4 be satisfied, however (e.g., Davidson 1985; McLaughlin 1988; Szabados 1985), and I have no objection to including 4 in a list of jointly *sufficient* conditions. Naturally, in some cases, whether the weight of a person's evidence lies on the side of *p* or of ~*p* (or equally supports each) is subject to legitimate disagreement.[18]

Return to the static puzzle. The primary assumption, again, is this: "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*." I have already argued that the assumption is false and I have attacked two related conceptual claims about self-deception: that all self-deceivers know or believe truly that ~p while (or before) causing themselves to believe that *p*, and that they simultaneously believe that ~*p* and believe that *p*. In many garden-variety instances of self-deception, the false belief that *p* is not preceded by the true belief that ~*p*, nor are the two beliefs held simultaneously. Rather, a desire-influenced treatment of data has the result both that the person does not acquire the true belief and that he or she does acquire (or retain) the false belief. One might worry that the puzzle emerges at some other level, but I have addressed that worry elsewhere and I set it aside here (Mele 1987a, pp. 129–30).

The conditions for self-deception that I have offered are conditions specifically for entering self-deception in *acquiring* a belief. However, as I mentioned, ordinary conceptions of the phenomenon allow people to enter self-deception in *retaining* a belief. Here is an illustration from Mele 1987a (pp. 131–32):

> Sam has believed for many years that his wife, Sally, would never have an affair. In the past, his evidence for this belief was quite good. Sally obviously adored him; she never displayed a sexual interest in another man; . . . she condemned extramarital sexual activity; she was secure, and happy with her family life; and so on. However, things recently began to change significantly. Sally is now

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

arriving home late from work on the average of two nights a week; she frequently finds excuses to leave the house alone after dinner; and Sam has been informed by a close friend that Sally has been seen in the company of a certain Mr. Jones at a theater and a local lounge. Nevertheless, Sam continues to believe that Sally would never have an affair. Unfortunately, he is wrong. Her relationship with Jones is by no means platonic.

In general, the stronger the perceived evidence one has against a proposition that one believes (or "against the belief," for short), the harder it is to retain the belief. Suppose Sam's evidence against his favored belief – that Sally is not having an affair – is not so strong as to render self-deception psychologically impossible and not so weak as to make an attribution of self-deception implausible. Each of the four types of data-manipulation I mentioned may occur in a case of this kind. Sam may positively misinterpret data, reasoning that if Sally were having an affair she would want to hide it and that her public meetings with Jones consequently indicate that she is *not* sexually involved with him. He may negatively misinterpret the data, and even (nonintentionally) recruit Sally in so doing by asking her for an "explanation" of the data or by suggesting for her approval some acceptable hypothesis about her conduct. Selective focusing may play an obvious role. And even selective evidence-gathering has a potential place in Sam's self-deception. He may set out to conduct an impartial investigation, but, owing to his desire that Sally not be having an affair, locate less accessible evidence for the desired state of affairs while overlooking some more readily attainable support for the contrary judgment.

Here again, garden-variety self-deception is explicable independently of the assumption that self-deceivers manipulate data with the intention of deceiving themselves, or with the intention of protecting a favored belief. Nor is there an explanatory need to suppose that at some point Sam both believes that p and believes that $\sim p$.

4. Conflicting beliefs and alleged empirical demonstrations of self-deception

I have argued that in various garden-variety examples, self-deceivers do not simultaneously have beliefs whose propositional contents are mutually contradictory ("conflicting beliefs," for short). This leaves it open, of course, that some self-deceivers do have such beliefs. A familiar defense of the claim that the self-deceived simultaneously have conflicting beliefs proceeds from the contention that they behave in conflicting ways. For example, it is alleged that although self-deceivers like Sam sincerely assure their friends that their spouses are faithful, they normally treat their spouses in ways manifesting distrust. This is an empirical matter on which I cannot pronounce. But suppose, for the sake of argument, that the empirical claim is true. Even then, we would lack sufficient grounds for holding that, in addition to believing that their spouses are not having affairs, these self-deceivers also believe, simultaneously, that their spouses are so engaged. After all, the supposed empirical fact can be accounted for on the alternative hypothesis that, while believing that their spouses are faithful, these self-deceivers also believe that there is a significant chance they are wrong about this. The mere suspicion that one's spouse is having an affair does not amount to a *belief* that he or she is so involved. And one may entertain suspicions that p while believing that $\sim p$.[19]

That said, it should be noted that some psychologists have offered putative empirical demonstrations of self-deception, on a conception of the phenomenon requiring that self-deceivers (at some point) simultaneously believe that p and believe that $\sim p$.[20] A brief look at some of this work will prove instructive.

Ruben Gur and Harold Sackeim propose the following statement of "necessary and sufficient" conditions for self-deception:

1. The individual holds two contradictory beliefs (p and not-p).
2. These two contradictory beliefs are held simultaneously.
3. The individual is not aware of holding one of the beliefs (p or not-p).
4. The act that determines which belief is and which belief is not subject to awareness is a motivated act (Sackeim & Gur 1978, p. 150; cf. Gur & Sackeim 1979; Sackeim & Gur 1985).

Their evidence for the occurrence of self-deception, thus defined, is provided by voice-recognition studies. In one type of experiment, subjects who wrongly state that a tape-recorded voice is not their own nevertheless show physiological responses (e.g., galvanic skin responses) that are correlated with voice recognition. "The self-report of the subject is used to determine that one particular belief is held," while "behavioral indices, measured while the self-report is made, are used to indicate whether a contradictory belief is also held" (Sackeim & Gur 1978, p. 173).

It is unclear, however, that the physiological responses are demonstrative of *belief* (Mele 1987b, p. 6).[21] In addition to believing that the voice is not their own (assuming the reports are sincere), do the subjects also *believe* that it is their own, or do they merely exhibit physiological responses that often accompany the belief that one is hearing one's own voice? Perhaps there is only a *sub-doxastic* (from *doxa:* belief) sensitivity in these cases. The threshold for physiological reaction to one's own voice may be lower than that for cognition (including unconscious belief) that the voice is one's own. Furthermore, another team of psychologists (Douglas & Gibbins 1983; cf. Gibbins & Douglas 1985) obtained similar results for subjects' reactions to voices of acquaintances. Thus, even if the physiological responses were indicative of belief, they would not establish that subjects hold conflicting beliefs. Perhaps subjects believe that the voice is not their own while also "believing" that it is a familiar voice.

George Quattrone and Amos Tversky, in an elegant study (1984), argue for the reality of self-deception satisfying Sackeim and Gur's conditions. The study offers considerable evidence that subjects required on two different occasions "to submerge their forearm into a chest of circulating cold water until they could no longer tolerate it" tried to shift their tolerance on the second trial, after being informed that increased tolerance of pain (or decreased tolerance, in another subgroup) indicated a healthy heart.[22] Most subjects denied having tried to do this, and Quattrone and Tversky argue that many of their subjects believed that they did not try to shift their tolerance while also believing that they did try to shift it. They argue, as well, that these subjects were unaware of holding the latter belief, the "lack

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

of awareness" being explained by their "desire to accept the diagnosis implied by their behavior" (p. 239).

Grant that many of the subjects tried to shift their tolerance in the second trial and that their attempts were motivated. Grant, as well, that most of the "deniers" *sincerely* denied having tried to do this. Even on the supposition that the deniers were aware of their motivation to shift their tolerance, does it follow that, in addition to believing that they did not "purposefully engage in the behavior to make a favorable diagnosis," these subjects also believed that they did do this, as Quattrone and Tversky claim? Does anything block the supposition that the deniers were effectively motivated to shift their tolerance without believing, at any level, that this is what they were doing? (My use of "without believing, at any level, that *p*" is elliptical for "without believing that *p* while being aware of holding the belief and without believing that *p* while not being aware of holding the belief.")

The study does not offer any direct evidence that the sincere deniers believed themselves to be trying to shift their tolerance. Nor is the assumption that they believed this required to explain their behavior. (The required belief for the purpose of behavior-explanation is a belief to the effect that a suitable change in one's tolerance on the second trial would constitute evidence of a healthy heart.) From the assumptions (1) that some motivation *M* that agents have for doing something *A* results in their doing *A* and (2) that they are aware that they have this motivation for doing *A*, it does not follow that they believe, consciously or otherwise, that they *are* doing *A* (in this case, purposely shifting their tolerance).[23] Nor, *a fortiori*, does it follow that they believe, consciously or otherwise, that they are doing *A* for reasons having to do with *M*. They may falsely believe that *M* has no influence whatever on their behavior, while not having the contrary belief.

The following case illustrates the latter point. Ann, who consciously desires her parents' love, believes they would love her if she were a successful lawyer. Consequently, she enrolls in law school. But Ann does not believe, at any level, that her desire for her parents' love is in any way responsible for her decision to enroll. She believes she is enrolling solely because of an independent desire to become a lawyer. Of course, I have simply *stipulated* that Ann lacks the belief in question. But my point is that this stipulation does not render the scenario incoherent. My claim about the sincere deniers in Quattrone and Tversky's study is that, similarly, there is no explanatory need to suppose they believe, at any level, that they are trying to shift their tolerance for diagnostic purposes, or even believe that they are trying to shift their tolerance at all. These subjects are motivated to generate favorable diagnostic evidence and they believe (to some degree) that a suitable change in their tolerance on the second trial would constitute such evidence. *But the motivation and belief can result in purposeful action independently of their believing, consciously or otherwise*, that they are "purposefully engaged in the behavior," or purposefully engaged in it "to make a favorable diagnosis."[24]

As Quattrone and Tversky's study indicates, people sometimes do not consciously recognize why they are doing what they are doing (e.g., why they are now reporting a certain pain-rating). Given that an *unconscious* recognition or belief that they are "purposefully engaged in the behavior," or purposefully engaged in it "to make a favorable diagnosis," in no way helps to account for what transpires in the case of the sincere deniers, why suppose that such recognition or belief is present? If one thought that normal adult human beings always recognize – at least at some level – what is motivating them to act as they do, one would opt for Quattrone and Tversky's dual belief hypothesis about the sincere deniers. But Quattrone and Tversky offer no defense of the general thesis just mentioned. In light of their results, a convincing defense of that thesis would demonstrate that whenever such adults do not consciously recognize what they are up to, they nevertheless correctly believe that they are up to *x*, albeit without being aware that they believe this. That is a tall order.

Quattrone and Tversky suspect that (many of) the sincere deniers are *self-deceived* in believing that they did not try to shift their tolerance. They adopt Sackeim and Gur's analysis of self-deception (1984, p. 239) and interpret their results accordingly. However, an interpretation of their data that avoids the dual belief assumption just criticized allows for self-deception on a less demanding conception. One can hold (1) that sincere deniers, due to a desire to live a long, healthy life, were motivated to believe that they had a healthy heart; (2) that this motivation (in conjunction with a belief that an upward/downward shift in tolerance would constitute evidence for the favored proposition) led them to try to shift their tolerance; and (3) that this motivation also led them to believe that they were not purposely shifting their tolerance (and not to believe the opposite). Their motivated false beliefs that they were not trying to alter their displayed tolerance can count as beliefs that they are self-deceived in holding without their *also* believing that they were attempting to do this.[25]

How did the subjects' motivation lead them to hold the false belief at issue? Quattrone and Tversky offer a plausible suggestion (p. 243): "The physiological mechanism of pain may have facilitated self-deception in this experiment. Most people believe that heart responses and pain thresholds are ordinarily not under an individual's voluntary control. This widespread belief would protect the assertion that the shift could not have been on purpose, for how does one 'pull the strings'?" And notice that a belief that one did not try to alter the amount of time one left one's hand in the water before reporting a pain-rating of "intolerable," one based (in part) upon a belief about ordinary uncontrollability of "heart responses and pain thresholds," need not be completely cold or unmotivated. Some subjects' motivation might render the "uncontrollability" belief very salient, for example, while also drawing attention away from internal cues that they were trying to shift their tolerance, including the intensity of the pain.

Like Quattrone and Tversky, biologist Robert Trivers (1985, pp. 416–17) endorses Gur and Sackeim's definition of self-deception. Trivers maintains that self-deception has "evolved . . . because natural selection favors ever subtler ways of deceiving others" (p. 282, cf. pp. 415–20). We recognize that "shifty eyes, sweaty palms, and croaky voices may indicate the stress that accompanies conscious knowledge of attempted deception. By becoming unconscious of its deception, the deceiver hides these signs from the observer. He or she can lie without the nervousness that accompanies deception" (pp. 415–16). Trivers's thesis cannot adequately be assessed here; but the point should be made that the thesis in no way depends for its plausibility upon self-deception's requiring the presence of conflicting beliefs. Self-deception that satisfies the set of sufficient

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

conditions I offered without satisfying the "dual belief" requirement is no less effective a tool for deceiving others. Trivers's proposal hinges on the idea that agents who do not consciously believe the truth (*p*) have an advantage over agents who do in getting others to believe the pertinent falsehood (~*p*): consciousness of the truth tends to manifest itself in ways that tip one's hand. But notice that an *unconscious* belief that *p* provides no help in this connection. Indeed, such a belief might generate tell-tale physiological signs of deception (recall the physiological manifestations of the alleged unconscious beliefs in Gur and Sackeim's studies). If unconscious true beliefs would make self-deceivers less subtle interpersonal deceivers than they would be without these beliefs, and if self-deception evolved because natural selection favors subtlety in the deception of others, better that it evolve on my model than on the dual belief model Trivers accepts.

In criticizing attempted empirical demonstrations of the existence of self-deception on Sackeim & Gur's model without producing empirical evidence that the subjects do *not* have "two contradictory beliefs," have I been unfair to the researchers? Recall the dialectical situation. The researchers claim that they have demonstrated the existence of self-deception on the model at issue. I have shown that they have not demonstrated this. The tests they use for the existence of "two contradictory beliefs" in their subjects are, for the reasons offered, inadequate. I have no wish to claim that it is impossible for an agent to believe that *p* while also believing that ~*p*.[26] My claim, to be substantiated further, is that there is no explanatory need to postulate such beliefs either in familiar cases of self-deception or in the alleged cases cited by these researchers and that plausible alternative explanations of the data may be generated by appealing to mechanisms and processes that are relatively well understood.

5. The dynamic puzzle

The central challenge posed by the dynamic puzzle sketched in section 2 calls for an explanation of the alleged occurrence of garden-variety instances of self-deception. If a prospective self-deceiver, *S*, has no strategy, how can *S* succeed? And if *S* does have a strategy, how can *S*'s attempt to carry it out fail to be self-undermining in garden-variety cases?

It may be granted that self-deception typically is *strategic* at least in the following sense: When people deceive themselves they at least normally do so by engaging in potentially self-deceptive behavior, including cognitive behavior of the kinds catalogued in section 3. Behavior of these kinds can be counted, in a broad sense of the term, as *strategic,* and the behavioral types may be viewed as *strategies* of self-deception. Such strategies divide broadly into two kinds, depending on their locus of operation. *Internal-biasing* strategies feature the manipulation of data that one already has. *Input-control* strategies feature one's controlling (to some degree) which data one acquires.[27] There are also *mixed* strategies, involving both internal biasing and input control.

Another set of distinctions 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 *with the intention of deceiving oneself* (e.g., intentionally focusing on data of a certain kind with the intention of deceiving oneself into believing that *p*). Many skeptical worries about the reality of self-deception are motivated partly by the assumption that 3 is characteristic of self-deception.

An important difference between 2 and 3 merits emphasis. Imagine a 12-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. 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 – to deceive herself into believing this hypothesis, or to cause herself to believe this. Intentional cognitive activities that contribute even in a relatively straightforward way to self-deception need not be guided by an intention to deceive oneself.[28]

For the record, I have defended a detailed account of intentions elsewhere (Mele 1992a, Chs. 7–13). Intentions, as I view them, are *executive attitudes toward plans,* in a technical sense of "plan" that, in the limiting case, treats an agent's mental representation of a prospective "basic" action like raising his arm as the plan-component of an intention to raise his arm. However, readers need not accept my view of intention to be persuaded by the arguments advanced here. It is enough that they understand intentions as belonging no less to the category "mental state" than beliefs and desires do and that they view intending to do something, *A,* as involving being *settled* (not necessarily irrevocably) upon *A*-ing, or upon trying to *A*.[29] Notice that one can have a desire (or motivation) to *A* without being at all settled upon *A*-ing. Desiring to take my daughter to the midnight dance while also desiring to take my son to the midnight movie, I need to make up my mind about what to do. But intending to take my daughter to the dance (and to make it up to my son later), my mind is made up. The "settledness" aspect of intentions is central to their "executive" nature, an issue examined in Mele 1992a.[30]

My resolution of the dynamic puzzle about self-deception is implicit in earlier sections. Such strategies of self-deception as positive and negative misinterpretation, selective attending, and selective evidence-gathering do not depend for their effectiveness upon agents' employing them with the intention of deceiving themselves. Even the operation of cold mechanisms whose functioning one does not direct can bias one's beliefs. When, under the right conditions, such mechanisms are primed by motivation and issue in motivated false beliefs, we have self-deception. Again, motivation can affect, among other things, the hypotheses that occur to one and the salience of those hypotheses and of data. For example, Don's motivational condition favors the hypothesis that his paper was wrongly rejected,

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

and Sid's favors hypotheses about Roz's behavior that are consistent with her being as fond of him as he is of her. In "testing" these hypotheses, these agents may accentuate supporting evidence and downplay, or even positively misinterpret, contrary data without intending to do that, and without intending to deceive themselves. Strategies of self-deception, in garden-variety cases of this kind, need not be rendered ineffective by agents' intentionally exercising them with the knowledge of what they are up to; for, in garden-variety cases, self-deceivers need not intend to deceive themselves, strategically or otherwise. Since we can understand how causal processes that issue in garden-variety instances of self-deception succeed without the agent's intentionally orchestrating the process, we avoid the other horn of the puzzle, as well.

6. Intentionally deceiving oneself

I have criticized the assumption that self-deception entails intentionally deceiving oneself and that it requires simultaneously possessing beliefs whose propositional contents are mutually contradictory, and I have tried to show how occurrences of garden-variety self-deception may be explained. I have not claimed that believing that *p* while also believing that ~*p* is conceptually or psychologically impossible. But I have not encountered a compelling illustration of that phenomenon in a case of self-deception. Some might suggest that illustrations may be found in the literature on multiple personality. However, that phenomenon, if it is a genuine one, raises thorny questions about the *self* in self-deception. In such alleged cases, do individuals deceive *themselves,* with the result that they believe that *p* while also believing that ~*p*? Or do we rather have interpersonal deception – or at any rate something more closely resembling that than self-deception?[31] These are questions for another occasion. They take us far from garden-variety self-deception.

Intentionally deceiving oneself, in contrast, is unproblematically possible. Hypothetical illustrations are easily constructed. 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," and counts 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).[32]

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 husband is not having an affair (or that she is not seriously ill, or that her child is not using drugs) in the face of strong evidence to the contrary. Why is that? Readers convinced that self-deception does not require the simultaneous presence of beliefs whose propositional contents are mutually contradictory will not seek an answer in the absence of such beliefs in Ike. 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 intentionally and consciously executes his plan for so doing; ordinary self-deceivers behave quite differently.[33]

This indicates that in attempting to construct hypothetical cases that are, at once, paradigmatic cases of self-deception and cases of agents intentionally deceiving themselves, one must 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 leaves 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 – normally, intentions hidden from us – are at work in ordinary self-deception.[34] Without denying that "hidden-intention" cases of self-deception are possible, a theorist should ask what evidence there may be (in the real world) that an intention to deceive oneself is at work in a paradigmatic case of self-deception. Are there data that can *only* – or *best* – be explained on the hypothesis that such an intention is operative?

Evidence that agents desirous of its being the case that *p* eventually come to believe that *p* owing to a biased treatment of data is sometimes regarded as supporting the claim that these agents intended to deceive themselves. The biasing apparently is sometimes relatively sophisticated purposeful behavior, and one may assume that such behavior must be guided by an intention. However, as I have argued, the sophisticated behavior in garden-variety examples of self-deception (e.g., Sam's case in sect. 3) may be accounted for on a less demanding hypothesis that does not require the agents to possess relevant intentions: for example, intentions to deceive themselves into believing that *p,* or to cause themselves to believe that *p,* or to promote their peace of mind by producing in themselves the belief that *p*. Once again, motivational states can prompt biased cognition of the sorts common in self-deception without the assistance of such intentions. In Sam's case, a powerful motivational attraction to the hypothesis that Sally is not having an affair – in the absence both of a strong desire to ascertain the truth of the matter and of conclusive evidence of Sally's infidelity – may prompt the line of reasoning described earlier and the other belief-protecting behavior. An explicit, or consciously held, intention to deceive himself in these ways into holding on to his belief in Sally's fidelity would undermine the project, and a hidden intention to deceive is not required to produce these activities.

Even if this is granted, it may be held that the supposition

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

that such intentions always or typically are at work in cases of self-deception is required to explain why a motivated biasing of data occurs in some situations but not in other very similar situations (Talbott 1995). Return to Don, who is self-deceived in believing that his article was wrongly rejected. At some point, while revising his article, Don may have wanted it to be true that the paper was ready for publication, that no further work was necessary. Given the backlog of work on his desk, he may have wanted that just as strongly as he later wanted it to be true that the paper was wrongly rejected. Furthermore, Don's evidential situation at these two times may have been very similar: for example, his evidence that the paper was ready may have been no weaker than his later evidence that the paper was wrongly rejected, and his evidence that the paper was not ready may have been no stronger than his later evidence that the paper was rightly rejected. Still, we may suppose, although Don deceived himself into believing that the article was wrongly rejected, he did not deceive himself into believing that the article was ready for publication: He kept working on it – searching for new objections to rebut, clarifying his prose, and so on – for another week. To account for the difference in the two situations, it may be claimed, we must suppose that in one situation Don *decided* to deceive himself (without being aware of this) whereas in the other he did not so decide; in deciding to do something, *A*, one forms an *intention* to *A*. If the execution of self-deceptive biasing strategies were a nonintended consequence of being in a motivational/evidential condition of a certain kind, the argument continues, then Don would either have engaged in such strategies on both occasions or on neither: again, to account for the difference in his cognitive behavior on the earlier and later occasions, we need to suppose that an intention to deceive himself was at work in one case and not in the other.

This argument is flawed. If on one of the two occasions Don decides (hence, intends) to deceive himself whereas on the other he does not, then, presumably, there is some difference in the two situations that accounts for *this* difference. But if there is a difference, *D*, in the two situations aside from the intention-difference that the argument alleges, an argument is needed for the claim that *D* itself cannot account for Don's self-deceptively biasing data in one situation and his not so doing in the other. Given that a difference in intention across situations (presence in one vs. absence in the other) requires some additional difference in the situations that would account for this difference, why should we suppose that there is no difference in the situations that can account for Don's biasing data in one and not in the other in a way that does not depend on his intending to deceive himself in one but not in the other? Why should we think that *intention* is involved in the explanation of the primary difference to be explained? Why cannot the primary difference be explained instead, for example, by Don's having a strong desire to avoid *mistakenly* believing the paper to be ready (or to avoid submitting a paper that is not yet ready) and his having at most a weak desire later to avoid mistakenly believing that the paper was wrongly rejected? Such a desire, in the former case, may block any tendency to bias data in a way supporting the hypothesis that the paper is ready for publication.[35]

At this point, proponents of the thesis that self-deception is intentional deception apparently need to rely on claims about the explanatory place of intention in self-deception itself, as opposed to its place in explaining differences across situations. Claims of that sort have already been evaluated here, and they have been found wanting.

Advocates of the view that self-deception is essentially (or normally) intentional may seek support in a distinction between self-deception and *wishful thinking*. They may claim that although wishful thinking does not require an intention to deceive oneself, self-deception differs from it precisely in being intentional. This may be interpreted either as stipulative linguistic legislation or as a substantive claim. On the former reading, a theorist is simply expressing a decision to reserve the expression "self-deception" for an actual or hypothetical phenomenon that requires an intention to deceive oneself or an intention to produce in oneself a certain belief. Such a theorist may proceed to inquire about the possibility of the phenomenon and about how occurrences of self-deception, in the stipulated sense, may be explained. On the latter reading, a theorist is advancing a substantive conceptual thesis: the thesis that *the* concepts (or our ordinary concepts) of wishful thinking and of self-deception differ along the lines mentioned.

I have already criticized the conceptual thesis about self-deception. A comment on wishful thinking is in order. If wishful thinking is not wishful *believing*, one difference between wishfully thinking that *p* and being self-deceived in believing that *p* is obvious. If, however, wishful thinking is wishful believing – in particular, motivationally biased, false believing – then, assuming that it does not overlap with self-deception (an assumption challenged in Mele 1987a, p. 135), the difference may lie in the relative strength of relevant evidence against the believed proposition: wishful thinkers may encounter weaker counter-evidence than self-deceivers (Szabados 1985, pp. 148–49). This difference requires a difference in *intention* only if the relative strength of the evidence against the propositions that self-deceivers believe is such as to require that their acquiring or retaining those beliefs depends upon their intending to do so, or upon their intending to deceive themselves. And this thesis about relative evidential strength, I have argued, is false.

Consciously executing an intention to deceive oneself is possible, as in Ike's case, but such cases are remote from paradigmatic examples of self-deception. Executing a "hidden" intention to deceive oneself is possible, too, but, as I have argued, there is no good reason to maintain that such intentions are at work in paradigmatic self-deception. Part of what I have argued, in effect, is that some theorists – philosophers and psychologists alike – have made self-deception more theoretically perplexing than it actually is by imposing upon the phenomena a problematic conception of self-deception.

7. Conclusion

Philosophers' conclusions tend to be terse; psychologists favor detailed summaries. Here I seek a mean. My aim in this paper has been to clarify the nature and relatively proximate etiology of self-deception. In sections 2–5, I resolved a pair of much-discussed puzzles about self-deception, advanced a plausible set of sufficient conditions for self-deception, and criticized empirical studies that allegedly demonstrate the existence of self-deception on a strict interpersonal model. In section 6, I argued that intentionally deceiving oneself is unproblematically pos-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

sible (as in Ike's case), but that representative unproblematic cases are remote from garden-variety instances of self-deception. Conceptual work on self-deception guided by the thought that the phenomenon must be largely isomorphic with stereotypical interpersonal deception has generated interesting conceptual puzzles. But, I have argued, it also has led us away from a proper understanding of self-deception. Stereotypical interpersonal deception is intentional deception; normal self-deception, I have argued, probably is not. If it were intentional, "hidden" intentions would be at work, and we lack good grounds for holding that such intentions are operative in self-deception. Furthermore, in stereotypical interpersonal deception, there is some time at which the deceiver believes that $\sim p$ and the deceived believes that p; but there is no good reason to hold, I have argued, that self-deceivers simultaneously believe that $\sim p$ and believe that p. Recognizing these points, we profitably seek an explanatory model for self-deception that diverges from models for the explanation of intentional conduct. I have not produced a full-blown model for this, but, unless I am deceived, I have pointed the way toward such a model – a model informed by empirical work on motivationally biased belief and by a proper appreciation of the point that motivated behavior is not coextensive with intended behavior.

I conclude with a challenge for readers inclined to think that there are cases of self-deception that fit the strict interpersonal model – cases in which the self-deceiver simultaneously believes that p and believes that $\sim p$. The challenge is simply stated: provide convincing evidence of the existence of such self-deception. The most influential empirical work on the topic has not met the challenge, as I have shown. Perhaps some readers can do better. However, if I am right, such cases will be *exceptional* instances of self-deception – not the norm.

ACKNOWLEDGMENT

Parts of this article derive from my "Two Paradoxes of Self-Deception" (presented at a 1993 conference on self-deception at Stanford). Drafts were presented at the University of Alabama, Université du Québec à Montréal, and Mount Holyoke College, where I received useful feedback. Initial work on this article occurred during my tenure of a 1992/93 NEH Fellowship for College Teachers, a 1992/93 Fellowship at the National Humanities Center, and an NEH grant for participation in a 1993 Summer Seminar, "Intention," at Stanford (Michael Bratman, director). For helpful written comments, I am grateful to George Ainslie, Kent Bach, David Bersoff, John Furedy, Stevan Harnad, Harold Sackeim, and BBS's anonymous referees.

NOTES

1. I have addressed many of these questions elsewhere. Mele (1987a) argues that proper explanations of both self-deception and irrational behavior involving *akrasia* or "weakness of will" are similar and generate serious problems for a standard philosophical approach to explaining purposive behavior. Mele (1992a) develops an account of the psychological springs of intentional action and the etiology of motivated rational and irrational behavior alike. Mele (1995) defends a view of self-control and its opposite that applies not only to overt action and belief but also to such things as higher-order reflection on personal values and principles; the book also displays the place of self-control in individual autonomy. Several *BBS* referees noted connections between ideas explored in this target article and those issues; some expressed a desire that I explicitly address them here. Although I take some steps in that direction, my primary concern is a proper understanding of self-deception itself. Given space constraints, I set aside questions about the *utility* of self-deception; but if my arguments succeed, they should illuminate the phenomenon whose utility is at issue. I also lack space to examine particular philosophical works on self-deception. On ground-breaking work by Audi (e.g., 1985), Bach (1981), Fingarette (1969), Rorty (e.g., 1980), and others, see Mele (1987b); on important work by Davidson (1982) and Pears (1984), see Mele (1987a, Ch. 10).

2. On the occasional rationality of self-deception, see Audi (1985; 1989) and Baron (1988, p. 40). On whether self-deception is an adaptive mechanism, see Taylor (1989) and essays in Lockard and Paulhus (1988).

3. For example, subjects instructed to conduct "symmetrical memory searches" are less likely than others to fall prey to the confirmation bias (see sect. 3). Subjects' confidence in their responses to "knowledge questions" is reduced when they are invited to provide grounds for doubting the correctness of those responses (Kunda 1990, pp. 494–95). Presumably, people who are aware of the confirmation bias can reduce their biased thinking by *giving themselves* the former instruction, and, fortunately, we do sometimes remind ourselves to consider both the pros *and* the cons before making up our minds about the truth of important propositions – even when we are tempted to do otherwise. For a review of the debate, see Kunda (1990). For a revolutionary view of the place of motivation in the etiology of beliefs, see Ainslie (1992).

4. Literature on the "paradoxes" of self-deception is reviewed in Mele (1987b).

5. One 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 a more modest partitioning view. For criticism of some partitioning views of self-deception, see Johnston (1988) and Mele (1987a, Ch. 10; 1987b, pp. 3–6).

6. This is not to say that self-deception is always "self-serving" in this way. See Mele (1987a, pp. 116–18) and Pears (1984, pp. 42–44). Sometimes we deceive ourselves into believing that p is true even though we would like p to be false.

7. Regarding the effects of motivation on time spent reading threatening information, see Baumeister & Cairns (1992).

8. The following descriptions derive from Mele (1987a, pp. 144–45).

9. For a challenge to studies of the vividness effect, see Taylor and Thompson (1982). They contend that research on the issue has been flawed in various ways, but that studies conducted in "situations that reflect the informational competition found in everyday life" might "show the existence of a strong vividness effect" (pp. 178–79).

10. This theme is developed in Mele (1987a, Ch. 10) in explaining the occurrence of self-deception. Kunda (1990) develops the same theme, paying particular attention to evidence that motivation sometimes primes the confirmation bias (cf. Silver et al. 1989, p. 222).

11. For a motivational interpretation of the confirmation bias, see Frey (1986, pp. 70–74).

12. See also Mele (1987a, pp. 125–26). See Bach (1981, pp. 358–61) on "rationalization" and "evasion," Baron (1988, pp. 258 and 275–76) on positive and negative misinterpretation and "selective exposure," and Greenwald (1988) on various kinds of "avoidance." Again, I am not suggesting that, in all cases, agents who are self-deceived in believing that p desire that p (see n. 6). For other routes to self-deception, including what is sometimes called "immersion," see (Mele 1987a, pp. 149–51, 157–58). On self-handicapping, another potential route to self-deception, see Higgins et al. (1990).

13. Literature on "selective exposure" is reviewed in Frey (1986). Frey defends the reality of *motivated* selective evidence-gathering, arguing that a host of data are best accommodated by a variant of Festinger's (1957; 1964) cognitive dissonance theory.

14. For references to work defending the view that self-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

deception typically is not intentional, see Mele (1987b, p. 11). See also Johnston 1988.

15. This is not to deny that self-deceivers sometimes believe that *p* while being aware that their evidence favors ~*p* (see Mele 1987a, Ch. 8 and pp. 135–36).

16. Condition 4 does not assert that the self-deceiver is *aware* of this.

17. On a relevant difference between being deceived *in* believing that *p* and being deceived *into* believing that *p*, see Mele (1987a, pp. 127–28).

18. Notice that not all instances of motivationally biased belief satisfy my set of sufficient conditions for self-deception. In some cases of such belief, what we believe happens to be true. In addition, since we are imperfect assessors of data, we might fail to notice that our data provide greater warrant for *p* than for ~*p* and end up believing that *p* as a result of a motivationally biased treatment of data.

19. This is true, of course, on "degree-of-belief" conceptions of belief, as well.

20. Notice that simultaneously believing that *p* and believing that ~*p* – that is, *Bp* & *B*~*p* – is distinguishable from believing the *conjunction* of the two propositions: *B*(*p* & ~*p*). We do not always put two and two together.

21. In a later paper, Sackeim (1988, pp. 161–62) grants this.

22. The study is described and criticized in greater detail in Mele (1987a, pp. 152–58). Parts of this section are based on that discussion.

23. For supporting argumentation, see Mele (1987a, pp. 153–56).

24. As this implies, in challenging the claim that the sincere deniers have the belief at issue, I am not challenging the popular idea that attempts are explained at least partly in terms of pertinent beliefs and desires.

25. Obviously, whether the subjects satisfy the conditions offered in section 3 as sufficient for self-deception depends on the relative strength of their evidence for the pertinent pair of propositions.

26. Locating such cases is not as easy as some might think. A *BBS* referee appealed to blindsight. There is evidence that some people who believe themselves to be blind can see (e.g., Weiskrantz 1986; cf. Campion et al. 1983). They perform much better (and in some cases, much worse) on certain tasks than they would if they were simply guessing, and steps are taken to ensure that they are not benefitting from any other sense. Suppose some sighted people in fact believe themselves to be blind. Do they also believe that they are *not blind*, or, for example, that they *see x*? If it were true that all sighted people (even those who believe themselves to be blind) believe themselves to be sighted, the answer would be *yes*. But precisely the evidence for blindsight is evidence against the truth of this universal proposition. The evidence indicates that, under certain conditions, people may see without believing that they are seeing. The same referee appealed to a more mundane case of the following sort. Ann set her watch a few minutes ahead to promote punctuality. Weeks later, when we ask her for the time, Ann looks at her watch and reports what she sees, "11:10." We then ask whether her watch is accurate. If she recalls having set it ahead, she might sincerely reply, "No, it's fast; it's actually a little earlier than 11:10." Now, at time *t*, when Ann says "11:10," does she both believe that it is 11:10 and believe that it is not 11:10? There are various alternative possibilities. Perhaps, although she has not forgotten setting her watch ahead, her memory of so doing is not salient for her at *t* and she does not infer at *t* that it is not 11:10; or perhaps she has adopted the strategy of *acting as if* her watch is accurate and does not actually *believe* any of its readings. (Defending a promising answer to the following question is left as an exercise for the reader: What would constitute convincing evidence that, at *t*, Ann believes that it is 11:10 and believes that it is not 11:10?)

27. Pears identifies what I have called internal biasing and input-control strategies and treats "acting as if something were so in order to generate the belief that it is so" as a third strategy (1984, p. 61). I examine "acting as if" in Mele (1987a, pp. 149–51, 157–58).

28. For further discussion of the difference between 2 and 3 and of cases of self-deception in which agents intentionally selectively focus on data supportive of a preferred hypothesis (e.g.) without intending to deceive themselves, see Mele (1987a, pp. 146, 149–51).

29. Readers who hold that intending is a matter of degree should note that the same may be said about being settled upon doing something.

30. For criticism of opposing conceptions of intention in the psychological literature, see Mele (1992a, Ch. 7). On connections between intention and intentional action, see Mele (1992a; 1992b; and Mele & Moser 1994).

31. Similar questions have been raised about partitioning hypotheses that fall short of postulating multiple personalities. For references, see Mele (1987b, p. 4); cf. Johnston (1988).

32. On "time-lag" scenarios of this general kind, see Davidson (1985, p. 145); McLaughlin (1988, pp. 31–33); Mele (1983, pp. 374–75; 1987a, pp. 132–34); Sackeim (1988, p. 156); and Sorensen (1985).

33. Some readers may be attracted to the view that although Ike deceives himself, this is not self-deception at all (cf. Davidson 1985, p. 145; McLaughlin 1988). Imagine that Ike had been embarrassed by his performance in class that day and consciously viewed the remark as ironic when he wrote it. Imagine also that Ike strongly desires to see himself as exceptionally intelligent and that this desire helps to explain, in a way psychotherapy might reveal to Ike, his writing the sentence. If, in this scenario, Ike later came to believe that he was brilliant in class that day on the basis of a subsequent reading of his diary, would such readers be more inclined to view the case as one of self-deception?

34. Pears (1991) reacts to the charge of incoherence, responding to Johnston (1988).

35. Talbott (1995) suggests that there are different preference rankings in the two kinds of case. (The preferences need not be objects of awareness, of course.) In cases of self-deception, the agents' highest relevant preference is that they believe "that *p* is true, if *p* is true"; and their second-highest preference is that they believe "that *p* is true, if *p* is false": Self-deceiving agents want to believe that *p* is true *whether or not* it is true. In the contrasting cases, agents have the same highest preference, but the self-deceiver's second-highest preference is the *lowest* preference of these agents: These agents have a higher-ranking preference "*not* to believe that *p*, if *p* is false." Suppose, for the sake of argument, that this diagnosis of the difference between the two kinds of case is correct. Why should we hold that in order to account for the target difference – namely, that in one case there is a motivated biasing of data and in the other there is not – we must suppose that an intention to deceive oneself (or to get oneself to believe that *p*) is at work in one case but not in the other? Given our understanding of various ways in which motivation can bias cognition in the absence of such an intention, we can understand how one preference ranking can do this while another does not. An agent with the second preference ranking may be strongly motivated to ascertain whether *p* is true or false; and that may block any tendency toward motivated biasing of relevant data. This would not be true of an agent with the first preference ranking.

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

If belief is a behavior, what controls it?

George Ainslie

Department of Veterans Affairs Medical Center, Coatesville, PA, 19320.
ainslie@coatesville.va.gov

Abstract: "Self-deception" usually occurs when a false belief would be more rewarding than an objective belief in the short run, but less rewarding in the long run. Given hyperbolic discounting of delayed events, people will be motivated in their long-range interest to create self-enforcing rules for testing reality, and in their long-range interest to evade these rules. Self-deception, then, resembles interpersonal deception in being an evasion of rules, but differs in being a product of intertemporal conflict.

Philosophical conundrums often serve to let us see flaws in our assumptions about ourselves – properties of our internal bookkeeping systems that prevent the books from balancing in particular circumstances. So it has been with Zeno's paradox, and so, I have argued elsewhere, with Newcomb's problem (Ainslie 1992, pp. 200–205) and with Kavka's problem (Ainslie 1994); and so it is, as Mele points out, with the seeming paradox of self-deception. That people hold beliefs categorically, in either/or fashion, and must abandon contradictory beliefs once they see the contradiction, is a useful norm for a philosopher, or lawyer, or perhaps a scientist; but the problem of self-deception demonstrates its failure as a candid description of how people think in every day life. Furthermore, the analogy to interpersonal deception calls for a reexamination of our assumptions about the unity of the self. Most important, the notion of purposive manipulation of one's own knowledge raises the question of how motivational determinants of beliefs interact with cognitive ones.

The assumption Mele criticizes is that only pathological beliefs are motivated by any consideration other than accuracy, or, perhaps, are motivated at all. However, the studies he cites (Gilovich 1991; Kunda 1987; Quattrone & Tversky 1984) (and he might have added Lewinsohn et al. 1980, and its descendents) demonstrate what people have always sensed in human nature: that to a large extent beliefs are cultivated or avoided insofar as they are occasions for pleasant or unpleasant feelings. When people "deceive themselves," they invariably seem to be discerning more occasion for good feeling (or less for bad) than a disinterested observer would, or than they themselves may discern in retrospect. Furthermore, this practice is usually thought to be a theft of good feeling from one's overall prospects, something that makes people worse off in the long run, if only because it makes them unrealistic and hence less effective at getting the very rewards for which they falsely hope. Such a view makes self-deception akin to addictions and other self-defeating behaviors, except that the studies just listed suggest that it is universally indulged in: Just as an alcoholic looks for excuses to go off the wagon, we all look for occasions to have good feelings sooner rather than later. The role of "rationality," then, is to discipline this impulse.

The above description does not offend intuition, but it does violate conventional utility theory. If a false belief impairs later reward-getting, the conventional theory implies a motive for people to cultivate it only if they steeply discount the future; and if they do discount the future steeply, or if the belief does not reduce future reward to begin with, conventional utility theory offers no motive for them to *avoid* this false belief. It has not been possible, therefore, to discuss in motivational terms what the constraints on belief might be. Such constraints are usually discussed as unmotivated, purely "cognitive" factors, even in areas like self-esteem maintenance and psychopathology, where the contingent feelings are intense (Perris 1988; Williams et al. 1988). Neither utility theory nor common sense has suggested ground rules for competition between the "motivation to be accurate" and the "movitation to arrive at particular conclusions" (Kunda 1990), a role for fact in a process of belief formation that seems ultimately to be under motivational control. If believing is a behavior – or depends on behavior – what factors induce self-deception and what factors limit it?

This is particularly puzzling for cases where self-deception has no practical cost, such as the belief that one is admired, or was the adopted child of a famous parent, or will one day come into an inheritance. Freud said that children give up living in fantasy because of the demands of reality (1915, p. 135), but in fact it is possible to live a protected life while thoroughly deluded, and a number of people do so. There must be something inadequate about fantasy *per se*, quite apart from its inadequacy as a means to practical ends, that leads most people to prefer the more rigorous demands of fact, and that enervates the emotional lives of those who dwell in fantasy.

Elsewhere I have proposed a mechanism for both the attraction of fantasy as belief, and its inadequacy for even purely aesthetic purposes (Ainslie 1992, pp. 243–63, 291–319). This mechanism is based on the discovery that both human subjects and animals spontaneously devalue future rewards according to a hyperbolic discount curve, rather than the exponential curve called for by conventional utility theory (Green et al. 1994; Kirby & Herrnstein 1995). The consequence of hyperbolic discounting is that temporary preferences tend to form for smaller, earlier rewards over larger, later ones, leading to a limited-warfare relationship – that is, a repeated prisoner's dilemma – among successive motivational states within the same individual.

This model supplies a mechanism for the competition between accuracy and wish fulfillment, *even within areas where beliefs are not instrumental to getting practical results.* Hyperbolic discounting predicts that those rewards that can be cultivated without the physical necessity of external stimuli, including the emotions that represent most of what rewards civilized people, will tend to be accessed in a pattern that selects early gratification over later but more intense gratification. It further predicts that the person will have a countervailing, weaker, but longer range interest to search for ways of forestalling this tendency. As the greater foresight of the latter interest takes effect, consumption strategies like fantasy, which access rewarding emotions *ad lib*, will gradually lose out. They will be dominated by strategies of accessing emotion according to cues that occur somewhat uncontrollably and even unpredictably, because the latter do not gratify the appetite for those emotions on demand and hence prematurely.

The events of external reality offer themselves as such independently generated cues, and the strategy of pacing emotion according to these cues (which, I argue, has the properties of belief), represents the discipline needed to prevent the dissipation of appetite by fantasy. Even beliefs that are obvious social constructions are not arbitrary; they must be constrained at least by this discipline (Ainslie 1993).

However, like all disciplines, adherence to reality is vulnerable to hedging, in this case, to alternative ways of collecting or interpreting observations that occasion better feelings in the immediate future, just as Mele describes. Such hedging is not an attempt to break down the discipline. Quite the contrary, it is an attempt to evade the discipline for only the time being, without substantially harming it. People who are tempted to deceive themselves want to defect in the prisoner's dilemma that they are playing with their selves at future times, but without causing these other selves to defect in turn.

Whether hedging on realism should be likened to interpersonal deception is probably a matter of taste. The attempt to defect in the present case without triggering a string of defections is apt to

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

include the information-distorting processes that Mele reviews. However, one's ability to manipulate the information available to oneself at another time is substantially more limited than one's ability to defraud another individual. The main issue in the intrapersonal case is the interpretation, not the awareness, of information. What is clear is that conventional assumptions about both beliefs and selves are inadequate to account for the data that are now available.

Self-deception vs. self-caused deception: A comment on Professor Mele

Robert Audi

Department of Philosophy, University of Nebraska, Lincoln, NB 68588-0321.
raudi@unlinfo.unl.edu

Abstract: Mele's study of philosophical and psychological theories of self-deception informatively links the conceptual and dynamic aspects of self-deception and explicates it without positing mutually inconsistent beliefs, such as those occurring in two-person deception. It is argued, however, that he does not do full justice to the dissociation characteristic of self-deception and does not sufficiently distinguish self-deception from self-caused deception.

Mele's wide-ranging, insightful study provides an excellent account of several philosophical and psychological theories of self-deception and the heretofore largely underexplored connections between them. He is particularly effective in linking the conceptual and dynamic aspects of self-deception. He does a real service to both philosphers and psychologists in arguing for a view of self-deception that explains the data without the paradox – or apparently insoluble problem – of positing flatly incompatible beliefs like those characterizing ordinary two-person deception, and without the puzzling presupposition that entering self-deception occurs through self-manipulative intentions. My aim here is to suggest why at one important point he may leave out too much of what the incompatible-belief view seeks to capture and why, correspondingly, he may construe self-deception too broadly.

1. Entering self-deception vs. getting oneself deceived. Consider Mele's proposed sufficient conditions for entering self-deception in the acquisition of a belief that *p* (sect. 3.2.4, para. 5). These conditions are typically sufficient, but suppose that, as a result of wanting to believe an airplane crash is due to mechanical failure, I seek to discuss the matter with people who believe this, and I thereby expose myself to one-sided evidence for it. This might cause me to speak to Eva, who turns out to believe a bomb was responsible and convinces me of that. This is compatible with my overall data favoring mechanical failure, which (let us suppose) is the true explanation. I thus have a false belief nondeviantly caused by treating relevant data in a motivationally biased way, as well as greater warrant for believing a contrary hypothesis. But is this self-decpetion? Granting that my actions produce my error, I may have been convinced by a good argument for a perfectly plausible hypothesis that happens to be false. True, a contrary hypothesis is better supported by my overall data; but suppose my data are complex and it is genuinely difficult to see which way they point. Would I then not simply be in a kind of error, one that would be possible even in highly rational persons and one not necessarily resulting from bias?

Mele might note that my motivationally biased handling of evidence is not a *direct* cause of my false belief. That is true, but the case might be revised to yield the same result directly (for some plausible notion of directness), as where a proper subset of my data do the same convincing that Eva's argument does.

My positive suggestion here is that what is missing (above all) is a certain *tension* that is ordinarily represented in self-deception by an avowal of *p* (or tendency to avow *p*) *coexisting* with knowledge or at least true belief that not-*p*. It is this tension that chiefly carries the crucial analogy to two-person deception that a plausible account of self-deception must preserve: just as when *A* deceives *B* in saying that *p*, typically *A* knows that not-*p* and *B* falsely believes *p*, so in self-deception *S* (sincerely) avows or is disposed to avow *p*, but at some level knows or truly believes not-*p*.[1] Since sincere avowal of *p* does not entail believing *p*, I can agree with Mele that self-deception does not require having incompatible beliefs; but because sincerely avowing *p* (or being disposed to avow it sincerely) is a main element in believing, this account captures something Mele is here omitting: the apparently dissociational phenomenon of sincere avowal – "virtual belief," one might almost say – together with knowledge that things are otherwise. In the airplane case, I am in no way dissociated. I simply have a false belief which results from trying to support my preferred hypothesis when in fact it is false; but if the data that convince me are plausible enough, I may be satisfied with, and settled in, my belief in a way that is at best rare with self-deceptive avowals.

2. Self-caused deception and ahistorical self-deception. My second concern also focuses on the breadth of Mele's view of self-deception. Recall the prankster who causes himself to believe a falsehood by entering it in his diary for a date at which he will have forgotten doing so. Mele grants this case will strike readers as "markedly dissimilar to garden-variety examples of self-deception," but he seems to allow it as an atypical case (at least if motivationally biased handling of evidence is an appropriate cause, as it may be). I suggest that this is not self-deception but only *self-caused deception* (see Mele 1997). Compare a case in which, for money, one does something designed to induce a false belief in oneself later. Even if, to facilitate the deed, one manipulates evidence to make the target proposition plausible, one can come to believe it so wholeheartedly that one's condition is indistinguishable from that of someone who simply has a false belief resting on *undivided* evidence that would seem adequate to any normal person.

One theoretical suggestion I am making, beyond the point that self-deception seems a kind of dissociational phenomenon, is that whether one enters it is determined more by the *kind* of state one enters than by the kind of *path* one takes in getting there. A familiar path can have a surprise termination; a phenomenon usually reached by a given process may sometimes be artificially induced in a way that would initially lead one to expect something different. These are not points Mele need deny. I emphasize them chiefly because they help put his examples in perspective.

The points also suggest an important question. Supposing that not all self-caused deception is self-deception: Must all self-deception be self-caused deception? Mele may seem to suggest a positive answer insofar as he implies that such behavioral manipulation of evidence are *the* route to self-deception; but I doubt he is committed to this by any major element in the target article. I would deny it on the (doubtless controversial) ground that self-deception is not a *historical* concept. If I am self-deceived, so is my perfect replica at the very moment of his creation.

If Mele's notion of self-deception is somewhat too broad, it should be remembered that he is at pains to describe the phenomenon of self-deception in a way that accommodates as many plausible accounts of that phenomenon as possible. Moreover, if I am right about the main reason why his conception is apparently too broad, that conception can be amended to deal with the problem with little change in essentials. He may still hold that there need be no paradox of self-deception, for instance, since incompatible beliefs are not required to understand it; it need not arise from intentional actions of putting oneself into a state of false belief; and self-deception may still be viewed, positively, as a phenomenon which manifests the effects of motivation on cognition.

NOTE

1. I have defended this idea in the papers cited by Mele and, most recently, in Audi (1997).

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Thinking and believing in self-deception

Kent Bach

Department of Philosophy, San Francisco State University, San Francisco, CA 94132. **kbach@sfsu.edu**

Abstract: Mele views self-deception as belief sustained by motivationally biased treatment of evidence. This view overlooks something essential, for it does not reckon with the fact that in self-deception the truth is dangerously close at hand and must be repeatedly suppressed. Self-deception is not so much a matter of what one positively believes as what one manages not to think.

Self-deception seems paradoxical if viewed on the model of deceiving someone else. This model is suggested by colloquial phrases like "fooling oneself" and "lying to oneself." Other philosophers wanting to avoid paradox (Bach 1981; Johnston 1988; McLaughlin 1988) have also rejected the assumption that underlie this model, that self-deception involves holding contradictory beliefs simultaneously and that it is engaged in this knowingly and intentionally. Putting these conceptual points to the empirical test, Mele shows in detail how recent psychological research can help make self-deception intelligible rather than paradoxical: the same processes involved in "cold biasing" can occur in self-deception, self-deception can be motivated without being intentional, and the self-deceiver's thinking can be purposeful without his being aware of what he is doing. Even so, Mele's deflationary account, as encapsulated by his proposed set of sufficient conditions on self-deception, leaves out something essential, something that, in my view, distinguishes self-deception from other sorts of motivated irrationality. As I see it, self-deception is not so much a matter of what one positively believes as what one avoids thinking.

Mele takes it as sufficient for self-deception that a person form or retain a false belief in the face of a preponderance of evidence to the contrary, doing so by treating his evidence in a motivationally biased way. This distinguishes self-deception from wishful thinking, which goes beyond one's evidence rather than conflicting with it outright, and from cold biasing, which is not motivated. However, it does not address the question of how the self-deceiver deals with the recurrent tendency of contrary thoughts to come to mind. Self-deception ordinarily involves more than a one-shot mistreatment of the evidence. It involves repeated avoidance of the truth, and this, I suggest, is not just a matter of belief.

For example, what makes the betrayed husband count as self-deceived is not merely that his belief that his wife is faithful is sustained by a motivationally biased treatment of his evidence. He could believe this even if he had no tendency to think about the subject ever again. He counts as a self-deceiver only because sustaining his belief that his wife is faithful requires an active effort to avoid thinking that she is not. In self-deception, unlike blindness or denial, the truth is dangerously close at hand. His would not be a case of self-deception if it hardly ever occurred to him that his wife might be playing around and if he did not appreciate the weight of the evidence, at least to some extent. If self-deception were just a matter of belief, then once the self-deceptive belief was formed, the issue would be settled for him; but in self-deception it is not. The self-deceiver is disposed to think the very thing he is motivated to avoid thinking, and this is the disposition he resists.

This view, that what matters is not the self-deceiver's belief but what he thinks (and does not think) when the touchy subject comes up, assumes a basic difference between thinking and believing. I take belief to be a complex of persistent dispositions concerning a certain proposition, whereas a thought is a relatively momentary though repeatable occurrence (Bach 1981, pp. 354–57). There are two kinds of disposition involved in belief, and only one of them is directly related to the occurrence of thoughts. This is the disposition to think the proposition one believes immediately when the subject comes up. If someone asks you the capital of Kentucky, your belief that it is Frankfort will lead you immediately to think that and answer accordingly. The other kind of disposition concerns a belief's role in cognitive processes, that is, to serve as a premise in reasoning and to limit the possibilities one considers in inquiry or in problem-solving. For example, one's belief that one never leaves one's eyeglasses in kitchen appliances contrains one's search for them – even if one does not actively think that one's glasses could not be in the refrigerator, one does not look for them there. Thus a belief can play a role in reasoning without actually coming to mind (Bach 1984). Even so, ordinarily a belief about something, when a subject relevant to it comes up, leads one to think the very thing one believes.

In self-deception, this tendency is inhibited. So even if the betrayed husband believes ("deep down," as we say) that his wife is unfaithful (I agree with Mele that such a belief is not necessary for self-deception), his otherwise normal tendency to think this is resisted. He might still act on this belief, for example, by regularly asking his wife where she has been, without explicitly thinking that she is unfaithful. To keep from think this he may need to clutter his mind with thoughts of contrary evidence (e.g., of his wife's displays of affection and words of assurance) but it is not necessary for self-deception that he actually believe that she is faithful. It is enough that he have thoughts of the sort that he would have if he did believe that and not have thoughts to the contrary, at least not on a sustained and recurrent basis.

How do self-deceivers manage to avoid a certain thought or at least rid themselves of it when it does occur? Three techniques have been distinguished (Bach 1981, pp. 357–62; Johnston 1988, p. 75; and McLaughlin 1988, pp. 51–55; the latter are discussed in Bach 1992): rationalizing (biased weighing of evidence), diverting one's attention (from where the evidence leads), and cluttering one's mind (with thoughts consistent with what one wants to be so). Self-deception can also be abetted by the self-serving use of what I call "exclusionary categories" (Bach 1994). Given our attentional and cognitive limitations, we must be selective in what we consider in a given situation and cannot spend time and effort on each thing that might come to mind just to determine that it is not worth considering. For this reason applying exclusionary categories, such as "absurd," "crazy," "impossible," and "irrelevant," can play a legitimate role in managing one's cognitive resources. In self-deception, though, such categories are applied in a motivationally biased way, thereby helping to keep an unpleasant truth from coming to mind.

Deceived by metaphor

John A. Barnden

Computing Research Laboratory and Computer Science Department, New Mexico State University, Las Cruces, NM 88003-8001.
jbarnden@crl.nmsu.edu; www.cs.nmsu.edu/jbarnden

Abstract: The views of self-deception that Mele attacks are thoroughly metaphorical, and should never have purported to imply the existence of real internal acts of deception. Research on self-deception, including Mele's appealing account, could be enriched and constrained by a broader investigation of the prevalent use of metaphor in thinking and talking about the mind.

Self-deception has interesting connections to metaphors of mind that are commonly used in everyday discourse. Papers on self-deception often use colorful metaphors, but such authors do not commonly *mention* the metaphors as such. I will suggest that the views of self-deception that Mele attacks are thoroughly metaphorical conceptions. Even though these conceptions may be pragmatically useful in common-sense thought and discourse about self-deception, it may well be that what is really occurring during self-deception is mental processing of the style claimed by Mele. Whether Mele is right or not, further study of self-deception would be enhanced by being placed in the context of an investigation of the use of metaphor in thinking and talking about the mind. (I have been engaged in such an investigation as part of an artificial intelligence research project – see Barnden [1996] and Barnden et al. [1995].)

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Ordinary speakers and writers frequently refer to parts of the minds of (mentally healthy) people as if those parts were themselves persons, with their own mental states, emotions, and so forth, and often engaging in natural language utterances. Here are some examples (from real discourse, with minor adaptations):

(1) "One part of Mike knows that Sally has left for good."
(2) "Part of Mike was insisting that Sally had left for good."
(3) "Part of you wants to talk about your personal problem but part of you hates the idea."
(4) "Half of me whispered that I'd drive all the way there."
(5) "It was as if his consciousness didn't want him to be without anxieties."
(6) "Did part of you think, 'Yes, I'm flattered'?"

I view such discourse chunks as manifestations of a conceptual metaphor, "mind parts as persons" (MPP). I would also claim that the metaphor is manifested in at least some readings of sentences such as

(7) "Sally told herself that Peter was faithful."

In many contexts as appealing, partial paraphrase would be that "one part of Sally was trying to convince another part that Peter was faithful."

In manifestations of the MPP metaphor, there is generally a strong connotation that mind parts that are *not* mentioned do *not* have the mentioned belief or desire. In (1) we should presumably take it that (metaphorically speaking) some other part of Mike does not know that Sally has left him for good. And often it is reasonable to take the unmentioned parts actually to have a contrary belief, desire, and so on. In (2) the use of "insisting" strongly suggests that (metaphorically speaking) some other part of Mike has claimed, and believes, that Sally had not left for good.

Now, the interpersonal view of self-deception that Mele attacks models self-deception on ordinary deception by one person of another. One common elaboration of this view is that the self-deceiving person contains two subsystems, one of which intentionally deceives the other into believing something. This elaborated view could, conceivably, be either a literal one or a metaphorical one, where of course a prime candidate for the metaphor is MPP. However, the literature on self-deception rarely mentions metaphor explicitly (but see some exceptions herewith), and discussions appear generally to assume without comment that the multiple-subsystems view is to be taken literally.

On the contrary, I suggest that the notion of self-deception, as it appears in stereotypical cases, is inherently metaphorical, and involves entirely metaphorical stances such as that one sub-person consciously intends to deceive another sub-person. It could well be useful and economical for us to adopt such a metaphorical view in everyday thought and discourse about self-deceivers. Nevertheless, the practical convenience of the metaphorical view does not imply that the objective, scientific truth of the matter is that the self-deceiver contains subsystems corresponding to the metaphorical sub-persons, or that, even if there are such sub-systems, that they do anything that could literally be called entertaining beliefs and intentions or engaging in acts of deceit, any more than the metaphor of death as a person implies that death really has beliefs and intentions. Therefore, the way is open for the mind to be operating in the way that Mele suggests it does in his appealing account.

Consider the question of what scientific sense can be made of "a part of" Mike believing something *P*, or "insisting" something, and so forth. In the case of such statements, where the use of metaphor is relatively blatant, it should not be considered shocking to claim that neither "Mike-as-a-whole" nor any identifiable sub-system within Mike can *literally* be viewed as believing *P*, insisting *P*, and so forth. Rather, we arguably have at most the right to say that *in some sense* Mike-as-a-whole believes *P*. And there is no contradiction in saying that Mike-as-a-whole in some sense believes *P* and in some sense believes *not-P*, because the "senses" could be different; we have no warrant to conclude that Mike-as-a-whole believes *P-and-not-P* in any sense.

Then, if I am right that common-sense notions of self-deception are inherently metaphorical, a scientific account of what really underlies self-deception should be continuous with a scientific account of what is going on behind garden-variety statements such as "a part of Mike believes." That being the case, there should be little impulse to suppose in the first place that self-deceit *really* involves a contradictory state of mind or any real intention by the agent or a subsystem of the agent to deceive anyone or anything. I therefore suggest that the study of self-deception could be broadened, enriched, and constrained by considering the relationships of that notion to other metaphorically described mental states and processes.

MPP is not the only relevant metaphor. In some accounts of self-deception, for example, that of Davidson (1985) alluded to in Note 5 in the target article, the self-deceiver's mind is viewed as partitioned into several regions, where the boundaries between regions cannot be crossed in some relevant sense: for example, the contradiction between *P* in one region and not-*P* in another cannot be seen by the person. These accounts rely on the metaphor of "mind as physical space," another extremely prevalent metaphor in ordinary discourse.

Finally, metaphor and related matters do occasionally receive mention in the literature on self-deception. For instance, Johnston (1988, p. 82) briefly mentions the tack of taking a subsystem account of self-deception metaphorically, but does not pursue the matter. Rorty (1988) bases an account on a superimposition of two "pictures" of the mind, and Bittner (1988, p. 538) casts talk of quasi-human parts of a self-deceiver as a "myth." Perhaps those authors would be happy to take the pictures and the myth, respectively, to be metaphors. Mele (1987, p. 3) quotes another self-deception researcher (King-Farlow 1963) as claiming that a person can quite often usefully be "looked at" as a large, loose sort of committee. Here King-Farlow is close to talking explicitly about MPP.

Biased steps toward reasonable conclusions: How self-deception remains hidden

Roy F. Baumeister and Karen Pezza Leith

Department of Psychology, Case Western Reserve University, Cleveland, OH 44106-7123. **rfb2@po.cwru.edu**

Abstract: How can self-deception avoid intention and conscious recognition? Nine processes of self-deception seem to involve biased links between plausible ideas. These processes allow self-deceivers to regard individual conclusions as fair and reasonable. Bias is only detected by comparing broad patterns, which individual self-deceivers will not do.

Mele has admirably exposed some fallacies in the common view of self-deception, particularly the notions of expecting simultaneous belief in both *p* and not-*p*, and of the intention to deceive oneself. In this comment, we extend Mele's analysis by showing how self-deception might operate. Most cases of self-deception involve beliefs about the self, and we propose that typical people hold multiple, contradictory views of what they may be like. Their best guess of the correct view may wander back and forth among those views, as social interactions yield encouraging or discouraging feedback.

Self-deception may operate in the processes by which people help their preferred views remain on top. As Mele says, there is neither the explicit intention nor the motivation to deceive oneself. The only operative motivation is the preference for the more attractive conclusion.

The crucial question, then, is how do people manage to avoid

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

recognizing that they deceive themselves? Mele says that people do not deliberately or intentionally engage in self-deception – but how can they pull it off without catching themselves in the act? One crucial explanation is that self-deception involves biased links between plausible ideas, so even careful scrutiny would only find each individual conclusion plausible. Self-deception is spotted only by comparing patterns of aggregated observations.

For example, Mele gives the example of a survey in which 25% of the respondents rated themselves as being in the top 1% in leadership ability. It is difficult to prove any single one of those claims wrong, and some of them were probably correct. It is only by virtue of aggregating them that we can say that self-deception must have been operating, insofar as the top 1% should only contain 1%.

To consider this argument systematically, we invoke the list of nine major self-deceptive techniques that one of us (Baumeister, in press) recently compiled from social psychology's studies of self-knowledge. The question for each case is whether individual claims could seem plausible, so that only by aggregating claims does the implausibility and hence self-deception emerge. Our central point is that the operation of prejudice is a matter of drawing biased conclusions while avoiding the appearance of bias, and this is accomplished by keeping each single cognitive step plausible and reasonable.

(1) The first pattern is the self-serving bias, by which people take greater responsibility for success than for failure. The difference is a matter of degree, and hence it is only detected by comparing self-attributions for success versus failure. By itself, each single responsibility judgment could be plausible.

(2) People discover more flaws in evidence that depicts them in an unflattering than a flattering light. The students who say the test was unfair are not the same ones who got the top grades. Again, though, the discrepancy that indicates self-deception is noticed only by comparing the groups. Many individual complaints may be valid.

(3) People spend less time processing unflattering feedback than flattering feedback. This, too, is a matter of degree that is only found by comparing large sets of responses. The amount of time any one person spent examining feedback would not by itself justify a charge of bias or self-deception.

(4) People remember successes and praise better than they remember failures and criticism. These biases may result in part from the different encoding times (and third point, above). It is doubtful that people try deliberately to forget bad news in most cases, and any single memory lapse would not be proof of self-deception (and especially not of intentional self-deception).

(5) People compare themselves to targets and standards they surpass. They can defend themselves against an accusation of bias by noting how accurately they made the comparison, thus totally avoiding the question of whether they chose an appropriate comparison target.

(6) People sort through their memories in biased ways to find evidence that they have desirable traits. These biases, too, can be seen only by contrasting searches done in different context. Each memory search does, after all, yield some facts that support the desired conclusion, and so everything the person remembers may be entirely correct and relevant.

(7) People overestimate the rarity of their good traits while at the same time thinking their faults and flaws are common. Meanwhile, they overestimate how many people share their opinions. Self-deception is implicit in these systematic distortions, but it requires the contrasting patterns to indicate that bias is at work. Each individual guesstimate may be quite reasonably and plausible.

(8) People shift the meaning of ambiguous traits so as to make themselves look good. In the earlier example of the leadership survey, different people may define leadership ability differently, especially in ways that help them convince themselves that they have it. In contrast, we doubt that 25% of men would rate themselves as being in the top 1% of height. Thus, ambiguity is again vital for successful self-deception, because it puts individual judgments beyond reproach.

(9) People in stigmatized groups can preserve their self-esteem by dismissing criticism as motivated by prejudice. Prejudice certainly exists and sometimes motivates criticism, and so no single response of this sort can be proven to involve self-deception. Indeed, it requires careful experimental evidence to show the pattern of bias on the part of recipients of criticism.

All in all, then, self-deception is not so much a matter of convincing oneself of proven falsehoods, as of steering of train of thought toward desired conclusions (Baumeister & Newman 1994). The processes are best hidden by a context of multiple, conflicting plausible conclusions and ambiguity. Like prejudice, self-deception may be difficult to prove in the single act and hence can be seen only after aggregating many responses and making suitable comparisons. Self-deceivers can thus assure themselves that their individual conclusions were reached by plausible, reasonable, justified inference processes.

Defending intentionalist accounts of self-deception

Jose Luis Bermudez

Department of Philosophy, University of Stirling, Stirling, FK9 4LA, Scotland.
jb10@stir.ac.uk

Abstract: This commentary defends intentionalist accounts of self-deception against Mele by arguing that: (1) viewing self-deception on the model of other-deception is not as paradoxical as Mele makes out; (2) the paradoxes are not entailed by the view that self-deception is intentional; and (3) there are two problems for Mele's theory that only an intentionalist theory can solve.

1. Self-deception and other-deception. There is something paradigmatically irrational about the idea of an agent simultaneously avowing two contradictory beliefs, but nothing like this need occur when self-deception is understood on the model of other-deception because the two beliefs could be *inferentially insulated*. Positing inferential insulation is not just an ad hoc manoeuvre to deal with the static paradox of self-deception (in the way that dividing the self into deceiver and deceived would be), because there are familiar computational reasons for denying that an agent's beliefs are all inferentially integrated (the limitations of memory search strategies, etc.). An account of self-deception can involve the simultaneous ascription of beliefs that *p* and that not-*p* without assuming that those two contradictory beliefs are simultaneously active in any way that make the contradiction explicit. A good way of explaining what an explicitly contradictory belief that not-*p* amounts to here would be that it is one which would make the agent's avowal that *p* insincere. Surely there is no conceptual confusion in suggesting that an agent might sincerely affirm *p* and yet have an inferentially insulated belief that not-*p*.

2. The intentionalist account and the paradoxes of self-deception. That static paradox is not entailed by the view that self-deception is intentional. That self-deception is intentional means simply that an agent successfully cause himself to believe that *p*, where *p* is false. The falsity of *p* need not feature in the content of his intention, nor need he have any beliefs about it. The relevant intention might simply be to cause himself to believe that *p* irrespective of its truth-value.

More complicated is the dynamic paradox, according to which an agent's intentional self-deception must be defeated by his knowledge of his intentions. Mele holds that the dynamic paradox does apply to intentionalist theories, and is avoidable only by a partitioning strategy or by his own anti-intentionalist strategy. There are two reasons, though, why we should not be convinced by this.

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

First, self-deception can be intentional without the relevant intention being to deceive oneself. If the intention is simply to cause oneself to believe that *p* irrespective of its truth-value, then the self-deceptive strategy is not *prima facie* self-defeating. Mele might argue that the paradox will still reappear, because one cannot believe *p* without believing that *p* is true and one cannot believe that *p* is true only because one believes that one has caused oneself to believe it. But this confuses the normative and the descriptive. No doubt one ought not believe that *p* only because one believes that one caused oneself to believe that *p*. But as a simple matter of psychological fact, people can reconcile those two beliefs. One might believe, for example, that although one initially set out to cause oneself to believe that *p*, the evidence in favour of *p* was so completely overwhelming that one would have come to believe that *p* regardless.

Second, it is highly implausible that doing something intentionally entails doing it knowingly (cf. Mele 1992a, p. 112), and nothing less than this will generate the dynamic paradox. Even the view (which entails the falsity of all Freudian accounts of repression) that when one is acting intentionally, what one is trying to do is accessible to introspection, will not generate the paradox because an intention can be accessible to introspection (in the sense that it could be brought to consciousness) without actually being conscious, and unless the intention is actually rather than potentially conscious, there is no reason for it to undermine the strategy of self-deception.

3. Two problems for Mele. According to Mele, the self-deceived acquisition of a belief that *p* requires only that a subject be nondeviantly caused to acquire the belief that *p* by a motivationally biased treatment of data relevant to assessing the truth-value of *p*. Against this the intentionalist holds that the subject must intend to cause himself to believe that *p* by biasing his cognitive processes because (a) he desires to believe that *p* and (b) he believes that the best way to achieve this is to bias his cognitive processes in the ways that Mele discussed. Here are two problems that militate against Mele and in favour of the intentionalist position.

4. The selectivity problem. Mele draws a false analogy between familiar examples of unintentional cold bias and motivationally primed hot bias. The point about instances of cold bias (like the availability heuristic) is that they are nonselective. Experimental work shows that, *irrespective of subject matter,* subjects have a tendency to be influenced by considerations of accessibility. Self-deception, however, is paradigmatically selective. Any explanation of a given instance of self-deception will need to explain why motivational bias occurred in *that* particular situation. But the desire that *p* should be the case is insufficient to motivate cognitive bias in favour of the belief that *p*. There are all sorts of situations in which, however strongly we desire it to be the case that *p,* we are in no way biased in favour of the belief that *p*. How are we to distinguish these from situations in which we desire *p* and are biased in favour of the belief that *p?* Mele, it seems to me, cannot answer this. The intentionalist can, however, by holding that a desire that one should believe that *p* is present in the latter but not the former situation.

5. The revision problem. It is perfectly conceivable that of two given individuals who believe that *p* and who desire equally strongly that it be the case that *p,* one should revise the belief that *p* in the face of given evidence whereas the other (self-deceivingly) refuses to accept that the evidence really is evidence against *p*. Presumably, Mele would say that the second is motivated by his desire for *p* to misinterpret negatively the *prima facie* evidence against *p*. But *ex hypothesi* the desire that *p* be the case cannot be all that is required to motivate the bias in question – otherwise both would be biased. So what motivates the bias in the second subject? Again, I do not see how Mele can answer this. The intentionalist can answer, though, by holding that the relevant difference between the two individuals is that only the second intends to cause himself to believe that *p*.

ACKNOWLEDGMENT

This commentary was written while I was a visiting scholar at Darmouth College.

Varieties of self-deception

Robert F. Bornstein

Department of Psychology, Gettysburg College, Gettysburg, PA 17325.
bbornste@gettysburg.edu

Abstract: Mele's analysis of self-deception is persuasive but it might also be useful to consider the varieties of self-deception that occur in real-world settings. Instances of self-deception can be classified along three dimensions: implicit versus explicit, motivated versus process-based, and public versus private. All three types of self-deception have implications for the scientific research enterprise.

Mele has made a persuasive argument that in its most elemental form self-deception is more straightforward and less complex than psychologists and philosophers have made it out to be. Stripped of its excess baggage (e.g., its connection to interpersonal deception in some analyses, and Freudian ego defenses in others) self-deception emerges as simpler – and less mysterious – than many of us thought. Mele's thoughtful analysis goes a long way toward placing self-deception squarely within the context of modern cognitive psychology.

Mele has simplified an unnecessarily complicated concept, but after simplification comes a different kind of complication, as researchers explore the varieties of self-deception that emerge in different situations and settings. Mele is certainly right in arguing that self-deception need not involve unconscious processes or defensive (i.e., motivated) distortion of internal and external reality. Yet even though it need not involve these things, in certain cases it probably does, and it is worth exploring the varieties of self-deception that take place in real-world settings. Instances of self-deception can be classified along three dimensions:

Implicit versus explicit. Certain instances of self-deception are implicit (i.e., involve processes that occur outside conscious awareness). Others are explicit – more deliberate, more conscious, and more controllable through traditional cognitive strategies. Just as researchers have found it useful to distinguish implicit memory from explicit memory, implicit learning from explicit learning, and implicit perception from explicit perception (see, e.g., Greenwald & Banaji 1995), it may be useful to distinguish implicit self-deception from explicit self-deception and to explore the contrasting dynamics of each.

Motivated versus process-based. Mele correctly notes that although self-deception can be motivated (e.g., by a desire to protect oneself from unpleasant information), it need not be. Perhaps we should distinguish motivated self-deception (arising from the kinds of self-protective mental activities described by Freudians and others) from process-based self deception, that is, self-deception inherent in the human information-processing apparatus (Kunda 1990). If internal need states and environmental variables differentially affect these two types of self-deception, this will allow researchers to explore the processes that underlie these two variants of self-deception in laboratory and field settings (Jacoby et al. 1992).

Private versus public. In some instances self-deception is a private act involving only the self. Alternatively, self-deception can involve both the self and others. The dynamics of private self-deception probably differ from the dynamics of public self-deception (although in the end this is an empirical question), just as private communications differ from those intended for multiple audiences (Fleming & Darley 1991). Public self-deception may be more cognitively effortful than private self-deception, and more easily disrupted by competing tasks and activities.

Whether it ultimately proves useful to examine different cate-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

gories of self-deception – or if instead a single unified model can account for these different instantiations – Mele's analysis of real self-deception remains compelling for two reasons. First, it informs us in important ways about everyday thinking, judgment, and reasoning. Second – and perhaps even more important – self-deception goes right to the heart of the scientific enterprise. As Mele notes, researcher's motivated to see their articles in print may well engage in self-deceptive thinking to achieve this goal. Other, equally salient instances of self-deception arise in science, and these, too, are worthy of consideration. For example, as I have argued elsewhere, manuscript reviewers who dislike manuscripts for reasons that cannot (or must not) be articulated publicly may deceive themselves into believing that the papers are unpublishable for some other, more acceptable reason (Bornstein 1991). The same is true of the journal editor forced to reject a sound manuscript for lack of journal space.

Following this line of thinking, one might argue that resistance to scientific progress is often rooted in self-deception, a desperate clinging to outmoded (but familiar) ideas in lieu of risky new concepts. In this context, our periodic Kuhnian revolutions may represent nothing more than a discipline-wide breakdown of shared self-deception in favor of a new, more compelling world view. Of course, when we recognize that ultimately the new paradigm will itself be supplanted by another, even more compelling framework, we must acknowledge that self-deception underlies not only resistance to scientific progress but scientific progress itself.

Paradoxical self-deception: Maybe not so paradoxical after all

Stephanie L. Brown and Douglas T. Kenrick

Department of Psychology, Arizona State University, Tempe, AZ 85287-1104. **asslb@asuvm.inre.asu.edu.**

Abstract: The simultaneous possession of conflicting beliefs is both possible and logical within current models of human cognition. Specifically, evidence of lateral inhibition and state-dependent memory suggests a means by which conflicting beliefs can coexist without requiring "mental exotica." We suggest that paradoxical self-deception enables the self-deceiver to store important information for use at a later time.

Is self-deception "irresolvably paradoxical"? We agree that in its "garden variety" it can be explained without stretching logic and without the assistance of "mental exotica." However, contrary to Mele's thesis, the *simultaneous* possession of logically contradictory beliefs can also be explained without any mysterious cognitive tricks. Holding conflicting beliefs about the same topic is not only possible, but eminently understandable within current models of human cognition.

Inconsistent beliefs and the modular brain. Mele's target article contains a tacit assumption that people have only one belief about a given topic at a given time. He discusses, for example, the distinction between a "mere suspicion" and a "belief," as if only one suspicion is eligible to graduate to the status of a belief. However, current research on the relationship between beliefs and behaviors indicates that people often have a number of different beliefs on the same topic (Cialdini et al. 1991). Which belief is activated depends on cues in the current context. When I am talking to my physician during my annual checkup I may fully believe alcohol has all the toxicity of strychnine. Later that same day, when I am chatting with my friends in a pub, I may just as fully believe that a few drops of the spirits can have all the benefits of ambrosia. Of course, given the 7-bit information processing limitation of working memory, only one of these beliefs is likely to be activated at a given moment. It is a mistake, however, to assume that the most recently activated belief somehow erases the others. Unless one engages in the intensive introspective housecleaning characteristic of psychoanalysts or philosophers, those contradictory beliefs on the same topic will continue to live side by side for many a year in long-term storage.

Current theories and research on human information processing support the possibility that the human mind can store apparently paradoxical information. For example, Martindale (1980; 1991) uses concepts such as lateral inhibition and state-dependent memory to explain the simultaneous existence of different "subselves" or executive systems within the same individual, each of which may have access to different memories. Through lateral inhibition, different processing units at the same level exert mutually inhibitory influences, thereby excluding all but one input to the next processing level. For example, lateral inhibition amongst letter recognition units allows us to perceive a given letter as either "a" or "d," but not both at the same time. Martindale argues that such processes operate at all levels in the nervous system, from single sensory features to features such as letters, words, sentences, and so on. At the highest level, Martindale argues that executive systems designated for centrally important tasks also inhibit one another, so that only one gains ascendancy at a given time. (The "subself" for reproductive behavior and that for surviving bodily threat are unlikely to be active at the same time, for example.)

Context-dependent memory provides one mechanism by which incompatible memories can exist within different executive systems. Via this memory process, it is easier to recall an episode or procedure in the same context in which it was learned (e.g., a name learned at a particular party with a particular blood-alcohol level). This phenomenon helps us understand how a belief acquired in one executive mode may exist simultaneously with a logically inconsistent belief acquired in another executive mode. For example, we may be led to believe that "free love" is a splendid idea while sexually aroused in the presence of an attractive partner, and to believe precisely the opposite after viewing a film about AIDs. In this sense, different parts of the self do exist simultaneously, but one wins the executive position at any given moment. If you get into an argument with your spouse, for instance, you easily retrieve examples of your spouse's malevolence to which you normally dedicate little processing time. On the other hand, during the pleasant making-up session you may well forget what you were arguing about earlier.

More recently, Michael Mills (1996) reviewed evidence suggesting that we replace the notion of one brain with that of "multiple brains," noting a "functional differentiation between several semi-independent macro-regulatory systems." Consistent with this type of argument, there is evidence that in different information-processing contexts we use very different types of logic (Tooby & Cosmides 1992). The concept of mutually independent executive systems, together with research on lateral inhibition and state-dependent memory, suggest that the storage of paradoxical information is quite possible, and requires no fantastic cognitive mechanisms.

When might self-deception make sense? From another perspective, it makes logical sense to suppress information that we may wish to store for use at a later time. For example, a man who, in most frames of mind, denies his wife is having an affair may nevertheless record the damaging bits of evidence for use when an alternative mate becomes available or when the benefits of remaining in the relationship decrease below some threshold. On the other hand, it often makes sense to evade explicit consideration of emotionally damaging information. The assumption that unpleasant memories could be stored and yet kept out of consciousness can be explained with concepts no more complex than those of classical conditioning, as Dollard and Miller (1950) long ago demonstrated.

Indeed, it may take effort above and beyond the normal call to work out all the logical inconsistencies among beliefs we have acquired in different frames of mind. Belief systems may be like works in progress – a potpourri of compatible, semi-contradictory,

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

and totally contradictory ideas. Unless we are writing a dissertation or engaging in psychoanalytic self-examination, we may rarely dedicate the effort required to pull all our beliefs together and come to a logical conclusion on most of what we think.

Once more with feeling: The role of emotion in self-deception

Tim Dalgleish

Medical Research Council Applied Psychology Unit, Cambridge, CB2 2EF, England. **tim.dalgleish@mrc-apu.cam.ac.uk; www.mrc-apu.cam.ac.uk/personal/tim dalgleish**

Abstract: In an analysis of the role of emotion in self-deception is presented. It is argued that instances of emotional self-deception unproblematically meet Mele's jointly sufficient criteria. It is further proposed that a consideration of different forms of mental representation allows the possibility of instances of self-deception in which contradictory beliefs (in the form p and $\sim p$) are held simultaneously with full awareness.

In his target article Mele does a convincing job, as he admits, of robbing the concept of self-deception of its mystery and paradox. As a motive, Mele confesses his admirable "desire to understand and explain the behavior of real human beings" (sect. 1, para. 4). Although I am largely in agreement with Mele's position, I will argue that his analysis dispenses not only with mystery and paradox but also with passion. I will further propose that an understanding of the role of *emotion* in instances of self-deception is as much about real human beings as the examples that Mele offers us.

I will endeavour to show that instances of emotional self-deception unproblematically meet Mele's jointly sufficient criteria (sect. 3.2.4, para. 5). However, this merely augments Mele's arguments concerning motivation. Less trivially, I will pick up the gauntlet thrown down in Mele's conclusion (sect. 7, para. 2) and argue that unexceptional cases of emotional self-deception can fit the strict interpersonal model. That is, they can involve holding two contradictory beliefs (p and $\sim p$) at the same time. Further, in contrast to Sackeim and Gur (Gur & Sackeim 1979; Sackeim & Gur 1978; 1985; sect. 4, para. 3) I will submit that it is possible and common for the self to be aware of *both* beliefs.

For the present purposes I use an Aristotelian or cognitive-functionalist model of emotions (e.g., Aristotle 1991; Lyons 1980; Oatley & Johnson-Laird 1987; Power & Dalgleish 1997). In this analysis, emotions are functional tools activated in accordance with certain cognitive appraisals of currently available information. Emotions have the function of consolidating that information and altering the circumstances that generated it. So, to take an example, information that a lion is running toward me will lead to an appraisal that the situation is threatening and the generation of the corresponding emotion of fear. Fear will function via reconfiguration of the relevant processing systems to collect as much evidence about the threat as possible, as quickly as possible, and to allow me to act promptly to reduce the threat, for example, by running away or shooting the lion.

If we take the above line, it is immediately clear that emotions can prime the various biasing mechanisms that Mele describes (sect. 3.1) in ways similar to the effects of motivations and desires (sect. 3.2). For example, Butler and Mathews (1983; 1987), in their research on the availability heuristic (sect. 3.1.2), showed that anxious individuals considered negative events (for example, experiencing a domestic fire) more likely to happen to themselves than to other people. In contrast, nonanxious individuals rated the events as equally likely to happen to themselves and others. In reality, anxious people are probably relatively less likely to experience such events because, on the whole, they tend to be more cautious than their nonanxious peers. However, a functionalist analysis proposes that anxiety primes the relevant processing systems in such a way that a disproportionate amount of threat-related, self-referent information relevant to the judgment at hand is "available" to anxious individuals. Consequently, their judgments concerning the self are biased. An important rider to this example is that it seems inappropriate to suggest that the anxious individuals in this study *desire* or are *motivated* to believe that they are in more danger than other people (see also Mele's Note 6). Rather, it is the emotional state they are experiencing that biases the relevant information-processing and leads to the self-deception.

As with Mele's examples involving motivation and desire (sect. 3.2), the above example of emotional self-deception concerns the acquisition of beliefs rather than their retention. However, garden-variety emotional examples can be found to illustrate the retention of self-deceptive belief structures. A particularly interesting case is that of jealousy as it contrasts with Mele's example of Sam believing that his wife is not having an affair (sect. 3.2.4, para. 11). In the case of jealousy it is common for the jealous person to selectively focus on and gather evidence in support of the partner's supposed infidelity even when the overwhelming weight of evidence supports the truth that there is no affair. Again, it is inappropriate to suggest that jealous persons desire or are motivated to find that their partners are unfaithful; rather, their emotional state is priming the relevant processing systems to gather evidence in a biased fashion.

The above instances of emotional self-deception are in line with Mele's deflationary analysis of the phenomenon. However, what about simultaneous possession of contradictory beliefs (p and $\sim p$)? One way to tackle this issue with respect to emotions is to explore the possibility of unconscious emotion-related beliefs that contradict consciously held views. However, Mele has little appetite for this approach as is clear from his brief discussion of hidden or Freudian intentions (sect. 6, para. 5). Consequently, I shall consider the more controversial case where both beliefs are available to conscious awareness.

The persuasiveness of any such line of argument hinges on the definition of belief. Mele reflects briefly on this issue in his discussion of the alleged empirical demonstrations of self-deception (sect. 4). Here he suggests that increased physiological responsiveness to certain information is not necessarily indicative of beliefs about that information. Indeed, he suggests that there is perhaps only a subdoxastic sensitivity in these cases. This view is clearly in line with multi-representational theories of mind (e.g., Multiple-Entry Memory Systems [MEMS] – Johnson & Multhaup 1992; Interacting Cognitive Subsystems [ICS] – Barnard 1985; Teasdale & Barnard 1993; Schematic, Propositional, Analogue, and Associative Representational Systems [SPAARS] – Power & Dalgleish, in press). In these approaches, various forms of nonpropositional representations are compatible with the subdoxastic process that Mele highlights. In addition, however, such multi-level theories describe higher-order meaning representations that are also nonpropositional and reflect constancies in lower level propositional and nonpropositional information.

Teasdale and Barnard illustrate the idea of two levels of meaning – a propositional and a higher-order, nonpropositional – by using the example of poetry. For example, "I wandered lonely as a cloud" and "I walked around aimlessly on my own" have the same meaning at the propositional level, but the understanding and the "sense" that each evokes is very different. The argument is that the line from Wordsworth activates higher-order meaning structures reflecting a sense of solitude and freedom that, it is suggested, is difficult to capture propositionally. The important point for the argument here is that these higher-order meaning structures are viewed as central to an understanding of emotions. An interesting proposal concerning self-deception follows from this; namely, that an individual can hold a propositional belief p while simultaneously having a higher-order emotional understanding of the situation consistent with $\sim p$. This mirrors the common phrase "knowing with the heart and knowing with the head." So, let us take the example of Ike's (sect. 6, para. 3) hypothetical brother, Mike. Mike may believe that his brother is an honest guy while at the same time having a sense that he is in fact deceitful.

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

There are a number of objections that Mele can make here. It could be argued that such higher-order representation of meaning, by virtue of being nonpropositional, would not constitute a belief (cf. the argument in sect. 4); that, in fact, this is some form of *supradoxastic* sensitivity. It could also be argued that such instances are remote from garden-variety examples. The former argument is more easily countered because people constantly translate higher-order semantic representations into propositional statements, albeit with some loss of meaning. For example, someone may say, "I know and believe that I am a success at work because I only have to look at the evidence, but deep down I still believe I am a failure." Regarding the remoteness of such examples, unless I am deceived, I would suggest that one need only monitor everyday conversation for a short time to realise that such paradoxical conflict is common. An even clearer (and perhaps increasingly garden-variety) demonstration is provided by any cursory content analysis of the transcripts of therapy sessions.

It may require another person to deceive oneself

Jean-Pierre Dupuy

CREA, Ecole Polytechnique, 75005 Paris, France.
jpdupuy@poly.polytechnique.fr

Abstract: There are other options than the opposition set by Mele between his own account and the strict interpersonal model. The notion of collective self-deception or "social hypocrisy" is discussed and shown to be nonparadoxical. When an individual consciousness lies to itself, there is often a form of "negative collaboration" with another.

Mele's deflationary account is cogent for the class of cases he considers, but not very thrilling. Lack of "conceptual fun" is something he claims as if to preempt criticism. However, his target article makes me think of mathematicians who would limit themselves to the case of two parallel lines while theorizing about conics: they would miss the concept of focus altogether. Statistics are irrelevant here. It may be that most instances of "real" self-deception are amenable to Mele's account. More often than not, what makes philosophy (and psychology) progress, nonetheless, is not the average but the exceptional.

Some measure of "intercultural exotica" may be appropriate to put things in broader prospective. Take common usage. One would be hard put to reach the conclusions Mele derives from dictionary definitions if one used the *Robert* instead of the *OED*. The extension of the French for "deceive," *tromper,* covers unintentional forms of deception to a much lesser degree than the English. Sartre's treating *mauvaise foi* (assuming this notion is germane to self-deception [Dupuy 1995]) as *lying* to oneself is characteristic, because lying implies the (hidden) intention to deceive (Sartre 1966). On the other hand, "unless I am deceived (i.e., mistaken)" translates as *Si je ne me trompe,* which literally means, "unless I am deceiving myself." What are we to make of these linguistic usages?

The challenge Mele issues in his conclusion may imprison us in the sterile alternative: either the simultaneous presence of contradictory beliefs, or Mele's account. Let me present another option, which I intentionally pick at an extreme of the broad spectrum covered by our (multilingual) usage of self-deception.

Consider two beings who are at once united and separated by a terrible secret. Think of those couples whose marriage has not been consummated after so many years because of the man's sexual impotence. The stability of such unions depends partially on the partners' never talking about *it*. Both of them know it, both of them know the other knows it, and so on, without, however, going to infinity. The opacity that results from this deviation from infinite reflexiveness allows one to pretend that the fact in question is not the case. This remains true so long as it is not *stated*, in other words, so long as it is not common knowledge (CK) (Aumann 1976; Lewis 1969).

Who is fooling whom here? This is not a case of lying or deceiving, and there is no irrational belief formation. One might speak of a kind of *negative collaboration* between two beings who accept, because it is convenient for them, a form of collective opacity. How can silence prevent CK? If a fact *F* is CK, everyone knows that *F* is CK. Hence the least doubt about CK's obtaining is proof that CK is not the case. As long as the man and the woman do not talk about *F,* the man has reason to believe that *F* is not CK. Therefore, *F* is not CK. One can go one step further and introduce a measure of individual self-deception along the lines of the interpersonal model. In one corner of his mind the man believes that *F* is CK, but in another corner he believes that *F* is not CK. Therefore, *F* is not CK.

If we go back to the other side of the Channel, we discover that this configuration, described by shared knowledge – everyone knows *F* – and an absence of CK, is treated as a particular form of lying to oneself. One might say the author and the victim of the lie in question are in this case the group itself. I am referring to the notion of social deception, or collective hypocrisy, which has played an essential role in French social sciences, as much in Durkheimian sociology as in the structuralism that has dethroned it. A famous example is the debate on "symbolic exchange." The question is: Do the natives know that behind the apparent generosity of gift exchange lies the sordid truth of economic interest and compulsory reciprocity? According to Bourdieu (1977), for example, the natives know the truth of reciprocity, but they hide it, for this truth is lethal to their social system. From whom do they hide it? From themselves, of course, themselves as a group. The problem is that the subject which is self-deceived is a nonsubject because it is the structure or the group.

Bourdieu's view becomes clearer, however, when he takes the example of a Kabyle worker who proclaimed the convertibility of the meal traditionally given at the end of work into money, with which he demanded to be paid instead. Bourdieu writes that this worker was only "*betraying the best-kept and the worst-kept secret: the one that is in everyone's keeping.*" This formula and others ("the secret is there is no secret"; "public secret," etc.) say nothing but what we described as a situation with shared knowledge but without CK.

Thus, the prima facie obscure notions of social hypocrisy or group self-deception prove nonparadoxical and amenable to a logical analysis of this kind. The detour via the collective, which could seem to introduce a formidable increase in complexity, has perhaps put us on the right track. What if, when an individual consciousness lies to itself, there were not this negative collaboration with another?

How many beliefs can dance in the head of the self-deceived?

Jeffrey E. Foss

Department of Philosophy, University of Victoria, Victoria, BC V8W 2Y2, Canada. **june19@uvvm.uvic.ca; jefffoss@uvic.ca**

Abstract: Mele desires to believe that the self-deceived have consistent beliefs. Beliefs are not observable, but are instead ascribed within an explanatory framework. Because explanatory cogency is the only criterion for belief attribution, Mele should carefully attend to the logic of belief-desire explanation. He does not, and the consistency of his own account as well as that of the self-deceived, are the victims.

Even if one day something should be found in the brain that might be identified with beliefs and desires, we cannot make this identification now. Because there is no way to observe them in themselves, beliefs and desires are, for all practical purposes, mere abstractions. Of course, abstractions are sometimes useful, even necessary, in explanations. We explain the orbit of the moon by

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

reference to its center of gravity, an imaginary point, and the pressure of a gas by its temperature, an abstract average of kinetic energy.

Explanations have a logic, and the logic of belief-desire explanations was outlined by Aristotle (though Foss 1976 is more accessible). Consider an example:

Jones believes that the glass contains deadly poison.
Jones desires that he not die.
So: Jones does not drink from the glass.

Aristotle noted that the same logic is involved in acting as in reasoning, and here the inference has the form of a *modus tollens:*

Jones believes (if I drink, then I die).
Jones desires (I do not die).
So, Jones acts (I do not drink).

In brief, an action is explained when its propositional content (the bit in parentheses) is entailed by the propositional contents of a belief *and* of a desire.

Alas, the subtleties of the human psyche often evade such simple logic, as in what we refer to as self-deception. The puzzles it generates arise solely from the breakdown of belief-desire explanation itself. Sadly, Mele's repairs merely generalize the breakdown.

First the problem itself. Suppose Al "deceives himself" about the cogency of his theory. In other words, Al is acting in ways that, *given* that we try to explain them using belief-desire psychology, lead us to attribute inconsistent beliefs to him. On one hand is his propounding of his theory and rebutting of criticism, which we explain by attributing to him the belief that his theory is cogent and the desire to believe what is true. On the other is his refusal to appreciate sympathetically the logic of critical commentary, which we may explain by attributing to him the belief that his theory is (or may be) unsound and the desire to avoid such a truth. We may amaze ourselves with the subtleties of the mind, that it might contain such tensions as contradictory beliefs and conflicting desires. But although it is logically impossible for contradictory propositions to both be true, there is no logical problem in their both being believed. If there is any mystery here, it is psychological rather than logical (Foss 1980).

Must we attribute contradictory beliefs to Al? Not at all. Beliefs and desires cannot be independently observed somewhere in the head, so the only constraint on their attribution is the cogency of the resulting explanation itself. Perhaps Al's failure to appreciate criticism stems not from his belief that his theory may be false, but from a single-minded conviction that his theory is cogent, coupled with his desire not to waste time. Or perhaps not. We may consult other aspects of Al's behavior. Perhaps a generally generous acceptance of criticism indicates confidence and self-transparency, or perhaps hypersensitivity to criticism indicates an inferiority complex and self-delusion. The evidence can be fairly persuasive one way or the other, but, finally, Al's behavior can be explained by ascribing to him various sets of beliefs, desires, and intentions.

Mele is convinced that people seldom if ever have contradictory beliefs, not even the self-deceived. His conviction flies in the teeth of manifest evidence that people are often complex and seldom of one mind, but let that pass. What cannot pass is his destruction of the very logic of belief-desire explanation that would give his zealous defense of human consistency its point. The key element of his defense is that "motivation sometimes biases beliefs" (sect. 3). "Motivation" is another word for desire, and the bias Mele has in mind, "a tendency to believe propositions we want to be true" (sect. 3). But wants and desires have no explanatory force without associated beliefs. The desire to drink will not *by itself* explain drinking. Unless it is also *believed* that there is something to drink – water rather than oil, something salutary rather than poisonous – the explanation is incomplete. Motivational states must be linked to information states to explain behavior. The desire to win explains every move in the chess game equally, but the different moves on each turn stem from different beliefs about the state of play at that turn.

The evidence Mele cites favoring the thesis that desire biases belief begs the question: he must explain why some desires that *p* lead to belief that *p*, while others do not. Al's desire that the theory he accepts be cogent might make him self-deceptively accept that it is, *or,* might make him suspicious of its cogency so he can avoid error. Belief-desire logic permits different actions to spring from the same desire given different beliefs. The desire that *p* would cause the belief that *p* only given some belief such as thinking something makes it so, or what you do not know cannot hurt you, but these beliefs will *contradict* beliefs such as that *p* is *really* true, and that one is not pretending that *p*. Given that consistency of belief for the self-deceived is his goal, Mele is not out of the woods yet.

If Mele feels compelled to defend consistency of belief among the self-deceived, I, for one, will not protest if he wants to modify the logic of belief-desire explanations. Unfortunately, he never even considers this logic. His assumption that the mere desire that *p* causes (or tends to cause) the belief that *p* makes us all out to be self-deceivers, in conflict with his original goal, whereas distinguishing self-deceivers from others requires attributing inconsistent beliefs to them. Either way, consistency of belief is a victim.

Self, awareness of self, and the illusion of control

Walter J. Freeman

Department of Molecular & Cell Biology, University of California at Berkeley, Berkeley, CA 94720-3200. **wfreeman@garnet.berkeley.edu; http://sulcus.berkeley.edu**

Abstract: A distinction between the self and its superstructure, the ego, supports Mele's conclusions. The dynamics of the limbic system generates the self through behavior that is subject to societal observation. The rest of the brain contributes awareness that, by ingenious back-dating and rationalization, gives the ultimate in self-deception: the illusion of control of the self by its own derivative.

Perception. An occurrence of self-deception is a discordance between the two modes of perception each of us has of ourselves and others' selves. One is an objective mode through observation of our conduct and its impact on others; the other is a private mode through awareness, which is verbally reported to others. These two modes can now be supplemented by direct observations of brain dynamics during the processes of perception. We are just beginning to realize the potential of this third mode for explaining features of consciousness previously inaccessible to us. One such feature is the delay in brain dynamics between choice and the consequent perception of choice. Who, or what, is in charge?

Materialists and cognitivists commonly view perception as a late stage of a process that begins with sensory transduction to form representations of stimuli, commonly in the firings of feature detector neurons. They hold that it proceeds through "binding" of the parallel activity of multiple features to represent objects, and then through the serial processes of normalizing, filtering, and matching with representations retrieved from storage for pattern completion and classification. They hold that perception is completed upon the binding of the representations of an object from the multiple sensory systems with an appropriate value or meaning attached to the fused image by the limbic system.

Existential brain dynamics. Studies of brain activity during perception by animals trained to discriminate olfactory, visual, auditory, or tactile stimuli (Barrie et al. 1996; Freeman 1975; 1992; 1995) have led to an alternative view, in which a percept is a goal-directed action that is organized by large-scale neural dynamics in the limbic system. Such action is intentional, because it forms

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

within a framework of space and time that has been constructed from the compendium of recent and remote actions and their sequellae (Piaget 1980), and because it is goal-oriented into the world (Freeman 1995).

This mode of brain function was put forth by the American pragmatists, most clearly by John Dewey (1914) in his critiques of the conditioned reflex. It was further developed by the Gestaltists, then by J. J. Gibson (1979, "affordances"), by Piaget (1980), and more recently by the situated actionists (Clancey 1993). Independently it was also developed in great detail by Merleau-Ponty (1942; 1945), who described it as "existential," in contradistinction from "empiricist" and "idealist" approaches. The originator of the existential view of brain function was Aquinas, who conceived the process of intentionality as the "stretching forth" by the brain through its body into the environment, and coming to know the environment through reshaping itself, what we now call learning through the plasticity of the brain (Freeman 1995).

The critical link between the public and private modes occurs at each moment of perceptual updating of the limbic pattern of activity, which incorporates an immediate result of intentional action as a fresh, small step along a trajectory extending into the future. The formulation of the next succeeding step is shaped by the entire body of past experience. With all its limitations of perspective and focus, this field of activity enacts choices and the awareness of choices, which is consciousness.

Discordance by delay. One key limitation is that time is required for neurodynamics to construct and reorganize the dynamic patterns following each definable stimulus. Libet (1994) has shown that there is a delay of 0.5 sec between the arrival of a stimulus and the onset of awareness of the stimulus, though that onset is subjectively back-dated to the actual time of arrival. Popper and Eccles (1977) describe this as having no physiological explanation. The process is analogous, however, to the two-threshold technique common only used by physiological psychologists for identifying that a response has occurred with a high threshold, and detecting when it has occurred with a preceding low threshold. Libet (1994) has extended his studies to show that similar delays occur in awareness of the initiation of intended actions.

Hence, the intentional, dynamic, public, limbic self continually constructs the neural activity patterns that instruct actions and seek sensory input. The global updating that sets the field for each next step lags by half a second. In this view the private experience of self, the "ego," is invariably half a second behind, always justifying, explaining, rationalizing, and claiming credit by virtue of back-dating, which was designed by evolution of the lemniscal system to keep the intentional self in synchrony with the unfolding real world. This is a cosmic joke on Descartes, whose vaunting ego got it backward. The existentialist says, "I am, therefore I think." Mahayana Buddhists and Lacanian psychoanalysts have written alike about the "illusion of the self." In the intentional view the illusion is not of the existence of the ego, but of the ego being in control of the self.

The intentional self can be observed by others in society as the seat of action, so it is assigned responsibility for action. It cannot be divided, except (according to Sperry, 1982) by splitting the brain, in contrast to the commonly splintered and bickering fragments of the private self-awareness. In this biologically based view, it is not surprising that an ego, if spinning off into a web of words and divorced from the testing afforded by bodily action, can, by habitual search of short-term gain (motivational bias) weaken its link to reality.

Is real self-deception really all that biased?

James Friedrich

Department of Psychology, Willamette University, Salem, OR 97301.
jfriedri@willamette.edu

Abstract: The mechanisms invoked to demonstrate how self-deception can occur without intention or awareness imply that self-deceptive beliefs are nevertheless the outcome of inappropriate and often egoistically driven processes. In contrast, models of pragmatic reasoning suggest that self-deception may well be the "reasonable" output of a more generalized, adaptive approach to hypothesis testing.

Mele does a great service by calling into question the conventional wisdom regarding self-deception, pointing out that exotic phenomena do not always require exotic explanations. He supports this position by outlining a variety of ways in which an actor might easily arrive at false beliefs without intending to self-deceive and without being aware of the inconsistencies inherent in the deception. In doing so, however, his implicit distinction between accuracy-driven and motivationally biased processing, and his relative emphasis on egoistic explanations, at times leads us perilously close to the kind of knowledgeable self-deception he seeks to discount.

Distinctions between accuracy motives and motives to confirm specific conclusions (e.g., Kunda 1990) invite us to treat self-deception as a distinct, qualitatively unique approach to hypothesis testing. Recent lines of theory and research, however, have suggested an alternative account of how hypothesis testing proceeds. These accounts are based on the notion that human cognitive systems are basically pragmatic in nature, well suited to minimizing costly mistakes and errors and only secondarily concerned with "truth detection" (Friedrich 1993; Lewicka 1989; 1992). In keeping with a Neyman-Pearson approach to judgment, the relative importance of false positive and false negative errors influences the appropriateness of strategies in various contexts.

For example, Lewicka (1989; 1992) reports that propositions involving positive/approach targets typically elicit sufficiently testing. Identifying sufficient conditions will regularly bring about desired outcomes, even if the resulting rule is too narrow and identifies conditions that are not necessary. In contrast, when targets are negative states or conditions, people appear to test appropriately for necessity, identifying conditions that – if absent or negated – allow one to avoid an aversive state. In a model similarly rooted in the adaptive, pragmatic qualities of reasoning, I have argued that lay hypothesis testing and data interpretation strategies are generally well suited to detecting and minimizing the errors most salient to an actor (Friedrich 1993). Indeed, strategies explicitly focused on truth detection can expose one to unnecessary risks. Such adaptive strategies probably operate in fairly automatic ways, with "normative accuracy" being more an occasional by-product of particular combinations of error concerns than an outgrowth of qualitatively distinct, accuracy-driven inference processes.

So what are the implications of such pragmatic strategies for understanding self-deception? In most empirical investigations of self-deception and biased information processing, there are relatively few costs associated with mistakenly giving oneself the benefit of the doubt (Friedrich 1993). According to normative strategies (but not to pragmatic reasoning strategies) such errors should be weighted equally with errors of self-denigration. For example, overestimates of personal control could conceivably have negative consequences, but these are rarely evident in the designs of studies demonstrating "illusions of control." Such overestimations of personal control may effectively serve to facilitate those behaviors that are necessary to increase the probabilities of desirable outcomes (Brown & Dutton 1995). Similarly, demonstrations of self-handicapping behavior rarely require subjects to engage in tasks where self-defecting behavior is particularly costly. Moreover, as Hirt et al. (1991) report, self-handicappers seem to prefer

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

to handicap themselves through distorted reports of emotional states rather than through explicit behaviors that would actually sabotage their performance.

But what about circumstances in which self-deception might have serious consequences? Mele brings out a central consideration when he discusses the example of suspected infidelity (sect. 4, para. 1). A person may "test" in ways that ultimately support a belief that one's partner is not having an affair. Yet this does not preclude the person's having suspicions and acting in ways that minimize the likelihood of disaster. By analogy, one can believe that nuclear power is extremely safe and yet one can behave in (costly) ways that acknowledge the risk and seek to fulfill the prophecy of safety. The malleability of errors is also important here; certain errors may be detectable yet perceived as unmodifiable. If so, there may be little pragmatic benefit to believing that one has certain negative characteristics of they are viewed as uncontrollable (Friedrich 1993). For example, if people were drawing inferences about their intelligence, they might show little concern for errors in which they falsely overlooked their low intelligence ("What could I do about it anyway?") and much greater concern for falsely accepting the validity of a test that had given them a low score (cf. Wyer & Frey 1983).

This is not to say that pragmatic strategies always lead to good conclusions. Rather, they are fairly efficient ways of detecting and minimizing salient errors: testing focused on both false positives and false negatives yields ostensibly normative behavior, and testing focused on the "wrong" kinds of errors can lead to disaster. Considerable evidence nevertheless suggests that garden variety self-deception may arise from very basic and adaptive cognitive processes that are not unique to self-relevant inferences. Mele takes an important step in showing how such processes make it unnecessary to assume that self-deception is a logical extension of interpersonal deception. Future work in this area may well go beyond this, serving to break down distinctions between truth-seeking and motivationally driven processes more generally. And such work will move us in the direction toward which Mele points us here, away from treating self-deception as a unique kind of cognitive processing and toward a model in which it is seen as an interesting manifestation of a more generalized, pragmatic testing strategy.

Detecting deception

Kenneth J. Gergen

Department of Psychology, Swarthmore College, Swarthmore, PA 19086.
kgergen1@swarthmore.edu

Abstract: I find three major shortcomings in Mele's account. First, verbal ambiguities suggest that the analysis is irrelevant to self-deception and/or that the traditional conception is subtly reinstated. Second, the data offer no means of establishing the superiority of the present account. Finally, as political rhetoric, Mele's proposal not only operates to disqualify others, but establishes science as their judge.

Having argued for the incoherence of the traditional view of self-deception (Gergen 1985), I can fully appreciate Mele's search for an adequate alternative. I am also stimulated by his proposal that the traditional conception rests on an account of interpersonal deception, reattributed to the domain of the psychological. Further, given the prevailing tendency of cognitive theorists to reduce the entirety of human function to cognitive universals, it is scarcely surprising to find the traditional concept of self-deception (lodged in psychodynamic theory) recast in the present manner. However, despite the carefully developed arguments and the range of evidence form which they draw, I must admit strong resistance to Mele's conclusions. Three particular problems bear articulation.

Linguistic mystification: Now you see it, now you don't. Are we treated here to a genuine advance in understanding or an academic shell game in which we are mystified by the subtle whisking of words? I fear the latter, with one deft maneuver simply eliminating the object, and the second, restoring it under a different shell. Consider the first: Mele effectively delineates the traditional (lexical) definition of self-deception, a definition pervasive both within the profession and society more generally. Appropriately finding it problematic, he then describes a series of studies on biased information processing. Such studies are used to support the conclusion, for example, that motivation can prompt cognitive behavior protective of favored beliefs, and that biased belief can function independently of motivation. It is then concluded that these studies demonstrate that self-deception can occur without its being intentional and without the individual harboring a relevant true belief. However, if there is no intention and no prior belief present in the act, then by traditional definitional standards, there is no self-deception. From the standpoint of common cultural sense, on what grounds should we consider these to be studies of self-deception at all? They were not seen as such by the investigators in question, their subjects would not see their activities in this way, and psychoanalysts would not view them as relevant to self-deception. In effect, as studies of cognitive bias, they have no obvious relevance to self-deception as commonly understood. We would face a similar case if a scientist announced that the culture fails to understand love because his explorations of sexual activity indicated a hormonal influence. The common meaning of the term is not illuminated but simply obliterated.

But then, in a second slight of words, aspects of the traditional meaning are subtly reasserted. Later in the target article Mele lays out four conditions sufficient for being self-deceived in a belief about *p*. The first is consistent with both his studies of biased information processing and traditional conceptions of self-deception, namely, that there is a belief in a false proposition. However, the second condition, motivational bias, operates in precisely the same way as intention in the traditional literature. Is there an empirically grounded difference between motivation and intention, or is this semantics at play? Putting aside the third condition (of nondeviant cases) as irrelevant to the issue at stake, we find a fourth entry into the arena of definition, not heretofore treated (at least in any forthright way). Here we learn that the body of data possessed by the person at the time would provide greater warrant for the conclusion that *p* is false. Yet, it is precisely some form of possession of the truth that is claimed by traditional self-deception theorists, and that Mele has been at pains to criticize in the target article up to this point. In effect, it appears that garden-variety self-deception has now reappeared in a different verbal guise.

Data terminal and interminable. Mele makes an interesting, and altogether apposite comment to the effect that we should consider self-deception as a theoretical construct that may or may not be adequate to explain the evidence at hand. Putting aside the challenging question as to whether there can be any evidence that is not already embedded in some form of theoretical understanding, we must also presume that the same argument holds for various information processing explanations. That is, the family of constructs favored by Mele (e.g., motivation, cognitive bias, priming,) are not facts in nature, but concepts that may or may not be useful in explaining the evidence. Given parity in explanatory potential, what is first unsettling about Mele's account is that the traditional self-deception theorist is granted no space for interpreting the studies described in the target article, and simultaneously no justification is offered or question raised concerning the adequacy of cognitive theory to explain the findings. Using Mele's definition, how are we to rule out the possibility that Mele himself is self-deceived in his conclusions?

More fundamentally, Mele's claim to explanatory superiority suggests some form of hermeneutic in which psychologists can successfully compare the explanatory capacities of their theories with respect to some form of psychological datum. The experimental data are, in effect, treated as if they are readings, manifestations, or expressions of some form of inner world to which theory should ultimately be responsive. As I have argued else-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

where (Gergen 1988), however, there is no way to distinguish among psychological mechanisms, processes, or the like, save through a theoretical *a priori*. Once initial agreements are secured concerning the mind and how it is manifest, theories can be compared; however, these agreements are in no way derived from observation of the events themselves. In effect, the mind may be viewed as a conversational object, and in the case of self-deception, a social construction of the professional psychologist (see also Gergen 1994).

Interpretation as politics by other means. Although Mele is clear enough about the role of theory as an explanatory device, there is otherwise an unfortunate tendency throughout the target article to reify the conceptual apparatus. Continuing a longstanding tradition in cognitive psychology, Mele comes to use terms such as self-deception, motivation, and the like as descriptions or stand-ins for the real. The very title of the piece, *Real self-deception*, is emblematic. Given the incapacity of theoretical language to picture or map the real, or to be linked ostensibly to particulars of the mind, how are we to respond to this invasive rhetoric of reality? At least one useful redoubt is to consider its cultural consequences. That is, the professional language of psychology is an entry into cultural life, and as this language is absorbed within its institutions and its daily relationships, we may be concerned with its consequences – ethical, ideological, and political. Here it is particularly worth noting that Mele's rendering of self-deception operates pragmatically in a highly similar way to the traditional account. That is, the term self-deception has traditionally operated as a performative, infirming and disqualifying the subject's avowals (see Gergen 1985). Although Mele attempts to redraw the conception, the pragmatic implications remain robust. In addition, this particular account thrusts the scientist (in this case the professional psychologist) into the role of arbiter on matters of self-deception. It is through scientific practice, we are subtly informed, that we rid ourselves of cognitive bias, and scientists themselves are positioned so as to rule on such matters. I worry about the unwarranted, unquestioned, and ultimately self-serving implications of the analysis.

There are alternatives. Many psychologists now seek means of theorizing the person in more relational terms (see Gergen 1994). That is, rather than viewing the individual as the site of rationality, motivation, and the like, the attempt is made to articulate the interpersonal matrix from which the human qualities of rationality, memory, and so on derive. Lewis (1996) nicely demonstrates the possibility of a relational analysis of self-deception, one that simultaneously places it within the sphere of human connection, and largely removes its pejorative implications. That seems a very promising direction for future work.

Partial belief as a solution to the logical problem of holding simultaneous, contrary beliefs in self-deception research

Keith Gibbins

Department of Psychology, Murdoch University, Murdoch, Western Australia, 6155 Australia. **gibbons@socs.murdoch.edu.au**

Abstract: A major worry in self-deception research has been the implication that people can hold a belief that something is true and false at the same time: a logical as well as a psychological impossibility. However, if beliefs are held with imperfect confidence, voluntary self-deception in the sense of seeking evidence to reject an unpleasant belief becomes entirely plausible and demonstrably real.

I agree entirely with the thrust of Mele's argument that there really is no such thing as self-deception, and with his general arguments, but I think another approach is equally effective in rejecting the idea. First I assume that, if the phenomena usually regarded as supporting the idea of self-deceiving behaviour can be accounted for even where the person is aware of the conflict between the two competing cognitions (e.g., being pro-Nazi and anti-Nazi, or seeing oneself as clever and as stupid about the same topic or problem), then people will have no problem with situations in which people cannot, after the event, report that there ever was a conflict: unconscious self-deception.

As I see it, the central puzzle Mele is attacking is whether intentional self-deception is logically possible and/or actually occurring. He quotes Gur and Sackeim (1979) as defining self-deception in terms of simultaneously holding a belief and its opposite (p and $\sim p$). It is this defining criterion I do not accept. Mele suggests in his caveat at the end of section 2 that he defies "believing p" as anything a person believes to a degree greater than 50%. Not-p ($\sim p$) is where the belief in p is less than 50%. Once we refuse to accept this pair of definitions the whole problem disappears. Partial belief simply states the common-sense idea that doubt exists.

If X is strongly motivated to believe one thing, but has strong doubts, that is, he believes it $<50\%$ – which is described by Mele as actually believing the opposite – he would be wise to look specifically for evidence designed to change his mind, that is, to increase his belief to $>50\%$. It is very hard to think of any situation in which there is no possible doubt whatever. The idea of uncertainty may in fact be totally general. Indeed, most philosophers warn us of the difficulty of even being absolutely sure that we have a table in front of us (when we do have one, that is!), and though they suggest that analytic statements are definitely true and so we have no reason to doubt them, anyone who has tried to add up a long column of figures or checked a computer program will know that the surety of truth in purely analytic systems does not, paradoxically, lead to any certainty that answers are correct.

The suggested "solution" is best presented by examples of the way doubts are suppressed and self-conversations adequately accomplished in the belief patterns of whole categories of people not just individuals. With the Inquisition on the alert, the sixteenth century ex-Jew who found Christian doctrine rather muddled and nonsensical, would be well motivated to find reasons to believe in it anyway. Similarly, any German living in Hitler's Germany around 1937 would be well aware that any doubts about Nazism that he previously held were safer being dismissed. In each case the person would be actively seeking to deceive in himself according to the definition Mele is using, but neither would be faced with any major logical problem. Each would be in a situation in which one says:

> I tend to believe this. It is dangerous to do so and I want to believe the opposite. I could be wrong. I hope I am wrong. Let me see if I cannot persuade myself that I am in fact wrong. Thank goodness, I have done it! I have changed my mind. Now let's make sure I do not have people or ideas coming along and persuading me I was right the first time. I will avoid the *possibility* by not listening to any arguments and avoiding people who believe what I used to believe. Or I could bravely persuade others of the rightness of my new views. If I can do that, I must have been right to change my mind or these converts would not be convertible, and I must admit sometimes I still have small doubts and need social support.

This idea of reinforcement of one's own faith by what amounts to missionary activity, predicts the keenness and fanaticism of the convert, and I have stolen it straight from Festinger et al. (1964).

In the attempt to persuade oneself, one could be expected to use every technique used when attempting to convert someone else if motivation were sufficiently great. We can assume that usually the motivation in self-deception is not so intense or at least not so clearly in one's best interests as in the chosen examples of the Gestapo and the Inquisition, in which case the persuasion effort might be less concentrated, but might nonetheless be very effective.

I believe the evidence suggests that the most common motive served by self-deception is self-esteem enhancement and protection, so perhaps it is not surprising that people do end up with views of themselves similar to those described in the quotation of

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Gilovich (1991) in section 3. Since the reviews by Jones (1973) and by Schrauger (1975) we have been well aware that people tend to accept information that flatters them provided there is little chance of having to come to terms with nasty reality. This certainly suggests a deliberate attempt to self-deceive, within Mele's definition, and to provide a more pleasant world view in which Self is better than expected, by seeking out and more readily accepting supporting evidence.

As indicated in Mele's various scenarios self-deception is often far from simple, but nothing in them seems to lead to important difficulties in handling the logical problems of self-deception.

Intentional self-deception can and does occur

Donald R. Gorassini

Department of Psychology, King's College, London, ON, N6A 2M3, Canada. **dgorassi@julian.uwo.ca**

Abstract: A form of self-deception exists that is both intentional and common. In it, people act as if they are undergoing a certain state of mind as a tactic for experiencing the state. This kind of self-deception can be illustrated by what happens to players of simulation games. Someone playing a pilot in a flight simulator game, for example, comes to experience aspects of the world of a pilot. Research on hypnotic responding is used to illustrate the nature and effectiveness of such a strategy of self-deception.

A form of self-deception exists that is both garden variety and more intentional than the type discussed by Professor Mele in the target article. I refer here to cases in which the person, in seeking to believe that a certain state (e.g., caring, anger, optimism) exists in the self, acts as if the state is occurring. In these cases, the person's knowledge that the state will not occur drives the fabrication of behavioral evidence designed to support the existence of the state. The (fabricated) evidence, in turn, helps convince the actor that the state is present. A good deal of theorizing in social psychology assumes that this intentional process of self-deception occurs in everyday life and is successful (e.g., Taylor 1989).

A look at what transpires with players of simulation games can help explain how intentional self-deception works (Gorassini, in press a). During play in a flight simulator – a sophisticated training and game technology – events can be organized perceptually by the player around one of two themes. One organizing framework is the reality defined by the game, in which the person is a pilot, the immediate surround is a cockpit, and the world beyond the plane's exterior is the sky. The other mode of organization consists of the reality defined by the situation that encompasses the game, in which the person is a player (not a pilot), the immediate environment is a fake cockpit (not a real one), and the area housing game apparatus is an arcade (not the sky). Human beings can control how they organize environmental input in simulator situations, much as they organize the stimulus input in so-called reversible figures (found in the perception chapters of introductory psychology texts). Events can be experienced in the game-defined way or experienced from the perspective encompassing the game. For extended periods, a player can get into the game and remain largely unaware that the game-defined theme is invalid. Self-deception in this model, then, is the perceptual shift from reality outside the game to reality inside the game followed by the extended use of game reality to define tasks to be performed.

The nature of the situation that the actor observes during a self-deception attempt is pivotal to the success of self-deception. If the flight simulator mimics well the sights, sounds, movements, and demands experienced in an actual aircraft cockpit, then self-deception has a much better chance of taking hold than if the simulator provides a poor representation on these stimuli. Several sources of realism exist in the simulator, including the appearance and actions of the principal actor, any supporting actors, and the nonhuman environment. This means that the player in the flight simulator must contribute to realism by assuming the role of a pilot. Failing to do so would make experiencing the world of a pilot flying an airplane impossible. If, all told, the information available to the actor provides a good counterfeit of game-defined reality, then self-deception becomes a relatively easy task.

In deceiving themselves, then, players carry out two kinds of intentional act, neither of which the Mele model of self-deception takes into account. The first is acting the role assigned by the game – caring person, competent person, or pilot. The second consists of construing events from the perspective defined by the game. The actor is spared the full burden of self-deception. A realistic game situation serves to help fool the self into believing that events are as they appear.

Research on hypnotic responding underscores the effectiveness of this process of self-deception. A response is hypnotic if it appears to occur involuntarily when suggested by the hypnotist. Research reveals interesting associated phenomena that suggest hypnotic responding is actually the product of an intentional self-deception process in which the person attempts to create the experiences, including involuntariness, that are thought to occur in hypnosis:

(1) Those who exhibit responses to hypnotic suggestions also frequently avow intentionally having made the response in an effort to experience hypnosis (Gorassini, in press b). This kind of report is suspiciously similar to the kind a game player would provide when describing what happened in a simulation game: "I acted like a pilot so I would feel like I was flying an airplane."

(2) Techniques designed to get research participants to interpret hypnosis as a game result in a substantial increase in the rate at which hypnotic responses are exhibited (Gorassini & Spanos 1986). This is even true of participants who previously scored low in responsiveness to suggestions. Because just about everyone possesses the ability to play simulation games and experience events as real within the game context, just about everyone can play the hypnosis game and feel, as a consequence, as if responses to suggestions are occurring involuntarily.

(3) Hypnotic responding and hypnosis-related experiencing occur most in situations made to appear prototypically hypnotic (Spanos 1986). When, for example, messages designed to elicit hypnotic responses imply the responses will be involuntary (e.g., "your arm is rising"), hypnotic responses, including experiences of nonvolition, occur more frequently than when the eliciting messages imply the responses will be cases of mudane obedience (e.g., "lift your arm") (Spanos & Gorassini 1984).

Using techniques such as role-playing, construal of events in terms of game reality, and the selection of situations known to support desired self-views, the actor intentionally self-deceives. Such a process is implied in social psychological theorizing in which it is believed commonplace for human beings to act their way into unwarranted beliefs about such things as their worth relative to others, their control over the environment, and the brightness of their future prospects (e.g., Swann 1987; Taylor 1989).

Self-deceived about self-deception: An evolutionary analysis

Mario Heilmann

Department of Psychology, University of California at Los Angeles, Los Angeles, CA 90095-1563. **mheilman@ucla.edu, mheilman@a3.com; www.a3.com/myself/**

Abstract: Mele's modified definition of self-deception is consistent with evolutionary theory. Self-deception is most likely whenever ignorance confers (reproductive) advantage, namely, in impression management, deception, conformity, social norms, reproductive knowledge, and existential conflicts. Second-order self-deception (unawareness of unawareness) perpetuates self-deception and may be the reason for our misguided definitions.

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

"Self-deception: A concept in search of a phenomenon," was the title of Gur and Sackeim's study in 1979. Seventeen years later, the search is still on. No phenomenon has been found to satisfy compellingly the "dual belief" criterion.

In contrast, the average person exhibits an abundance of empirically demonstrated everyday phenomena that fulfill Mele's more relaxed criteria of self-deception. In a review of the research, Taylor and Brown (1988) show that "overly positive self-evaluations, exaggerated perceptions of control or mastery, and unrealistic optimism are characteristics of normal human thought" (p. 193; see Taylor 1989 for more details). In contrast, mildly depressed people tend to be more realistic (pp. 196–197).

Yet, we do not tell the depressed that they need to get rid of realistic unbiased thinking and acquire positive illusions like everybody else. Much is at stake if we become aware of our self-deception. Research that would create awareness of our self-deception may even be dangerous to our psychological well-being. People with low levels of self-deception have higher levels of depression (Taylor & Brown 1988, p. 197) and other psychopathologies (Paulhus 1986).

Overconfidence, a tendency to express unwarranted subjective certainty (Baumann et al. 1991; Griffin 1990) and second-order self-deception, an unawareness of our unawareness, are essential elements. Details about censorship must be censored, otherwise our whole self-deceptive house of cards might fall into pieces, just as the Soviet Union did after censorship was lifted. Self-deception about self-deception is not just a game of words, it is an essential mechanism that has helped us keep our self-deception intact through the millennia.

Evolution. Evolutionary biology posits that we evolved mental mechanisms that helped us maximize our inclusive fitness in Pleistocene environments. Inclusive fitness is roughly defined as the number of copies of our genes in future generations' gene pool transmitted through our own offspring, and, to a lesser amount, through offspring of close genetic relatives.

Evolutionary analysis could enlighten the discussion about motives and strategies. No matter how motivated one is to jump off a cliff and fly, self-deception about this capacity would hardly increase our survival and reproductive success. Conversely, self-deception about one's below-average aptitudes may increase one's job or marriage prospects. Overconfidence, for example, makes physicians appear more secure and knowledgeable, and it thus increases patient satisfaction (Baumann et al. 1991, p. 167). Impression management and self-presentation, as described by Goffman (1959), are prime candidates for self-deception. It pays to look honest, like a good ally or desirable faithful mate, while avoiding paying the full reproductive cost entailed by actually behaving in such an altruistic manner. Hypocrisy and deception often confer (reproductive) advantage. In animals, deception is ubiquitous (Mitchell & Thompson 1986) and deceptive mimicry is often built into an animal's physiology. [See Whiten & Byrne: "Tactical Deception in Primates" *BBS* 11(2) 1988.]

Burley (1979, p. 844) suggests that concealed hominid ovulation is a built-in self-deceptive mechanism "to counter a human or pre-human conscious tendency among females to avoid conception through abstinence from intercourse near ovulation" (see Miller 1996 for a discussion of competing theories). Comprehension of procreative mechanisms is an unintended side effect of increased human intelligence. So is existential anxiety. Animals need no self-deception to defend against it, because healthy animals probably do not worry about death and afterlife.

Intelligence can also cause problems when applied to social rules and norms. Theoretically, we have sufficient brainpower to question the usefulness of painful tribal initiation rites, of genital mutilation, of compulsory use of veils, of suit and tie in tropical climates, or of unhealthy high-heeled shoes. Only if we are not part of the respective culture do we find such practices strange. The few individuals that do not accept the norms and socially created reality of their own culture suffer unpleasant sanctions.

Drinking carrot juice or milk instead of beer at a fraternity party might slightly increase one's life span, but would severely decrease popularity and dating success, and consequently, reproductive success. Wearing safety belts or being a nonsmoker when this was still considered uncool would similarly have reduced social success. Self-deceived individuals who selectively blank out their logical analysis when it conflicts with societal beliefs are frequently at an advantage.

Conclusions. It is quite likely that dual belief self-deception actually exists. Yet, by de-emphasizing this definition we can focus on the essential qualities and functions of self-deception. The over stringent definition of self-deception might be an unconscious attempt to muddy the waters and obscure the ubiquity of self-deception. It would be interesting to investigate how we, laypeople and experts alike, fail to face the truth about our own biases, illusions, and self-deception.

Real ascriptions of self-deception are fallible moral judgments

Edward A. Johnson

Department of Psychology, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada. **ed__johnson@umanitoba.ca**

Abstract: Mele's jointly sufficient conditions for self-deception preclude definitive ascriptions of self-deception in practice. Consequently, actual ascriptions of self-deception require large inferences and may frequently be in error. It is recommended that attention be directed toward actual practices of ascription to understand how children learn and adults dispense what is ultimately a moral judgment.

As a *theoretical* account of self-deception, I am persuaded that Mele has provided us with an essentially correct, viable model of the phenomenon. Rather than quibble, I wish to direct my remarks to some interesting problems that emerge only when applying his sufficient conditions for self-deception to actual cases. My argument is that Mele's conditions require knowledge on the part of would-be attributers of self-deception that they do not or cannot possess, thereby making definitive ascriptions of self-deception impossible. Let me illustrate.

Regarding Condition 1 (sect. 3.2.4), namely, that the belief acquired by S (p) is false, the problem here is that just as S does not know p is false, neither may anyone else. Although p may be false, and S should know it to be so, it is often the case that S is the only one who is in a position to know that p is false.

Regarding Condition 2, that S treat data that are "seemingly relevant to truth value of p in a motivationally biased way" (3.2.4), establishing what evidence is relevant to the truth of a proposition is a notoriously difficult task. One need not be motivationally biased to err in determining the relevance of evidence. Also, the requirement that the data be handled in a "motivationally biased way" seems to presuppose a standard of unmotivated or at least unbiased handling of evidence that does not exist. How, for example, does one determine what would constitute an unbiased weighting of conflicting pieces of evidence? Equal weighting is not necessarily any less biased than unequal weighting. Nothing about the actual weightings can tell us whether motivated bias is present.

Condition 4 requires that the body of data possessed by S at the time provide a greater warrant for $\sim p$ than p. Aside from the impossibility of ever knowing with certainty just what S knows or is aware of, the problem of weighting evidence arises again. For rational reasons one may be able to adjust the weights concerning the available evidence such that it provides a greater warrant for $\sim p$ than p.

Conclusion: definitive ascriptions of self-deception in everyday life are simply not to be had. Mele may justly complain that he never claimed they could be. However, I want to indicate now why this is of some consequence in any case.

First, the difficulties that prevent us from making definitive ascriptions of self-deception may well lie in the phenomenon

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

rather than the criteria proposed by Mele. One can argue on psychological grounds that entering or maintaining a state of self-deception depends on the relative inaccessibility of definitive grounds for knowing the truth about *p*. If true, then self-deception constitutes a kind of shadow-land epistemic phenomenon that comes into being only under cover of ambiguity or uncertainty, and evaporates when exposed to the light of overwhelming fact. Mele suggests as much when he observes that the evidence in instances of self-deception cannot be "so strong as to render self-deception psychologically impossible and not so weak as to make an attribution of self-deception implausible" (sect. 3.2.4). Thus, the epistemic conditions that make self-deception possible are the same ones that make its definitive ascription impossible. (Except, one might suppose, in retrospect, when the falseness of the belief is fully exposed, yet even here there are inferences to be made about what was known when.)

It follows, therefore, that ascriptions of self-deception in real life require inference, and hence their accuracy will vary with the sensitivity or bias of the observer. This fact raises interesting questions. Do observers actually use criteria anything like Mele's when ascribing self-deception? If they do, are there nonetheless differences in the thresholds they use? Conceivably, psychoanalysts might have lower thresholds than lay persons for attributing self-deception in given cases. Is this the result of greater acumen or only a less stringent threshold?

The possibility that biases and prejudices may influence ascriptions of self-deception becomes more of a matter for concern when it is recognized that such ascriptions are moral judgments (Schmitt 1988). That is, if S causes harm to another because of a false belief that he has deceived himself into accepting, then S cannot evade responsibility for the harm, because S ought to have known better than to accept the belief in the first place. In ascribing self-deception we are dispensing blame, not for bad actions but for faulty thinking.

Seen in this light, attributions of self-deception constitute a cultural practice about which we know little. What conception of the mind does it presuppose? Presumably, that the observer has an understanding of false belief, how evidence causes belief, and what one's responsibility is for engaging in even-handed epistemic practices. Developmentalists may therefore be interested to know how and when children learn to attribute self-deception. Research suggests that children come to understand the culpability of self-deceivers for their false beliefs at about age 8 or 9, at the end of a developmental sequence that progresses from an understanding of their own and others' false beliefs to the blameworthiness of other-deception, and culminates in an appreciation of the blameworthiness of self-deception (Johnson 1996).

To conclude, Mele has given us a means of understanding how self-deception is theoretically and psychologically possible. However, a full understanding of *real* self-deception must also encompass the practice of its ascription, including the developmental, motivational, and epistemological circumstances of the ascribers.

Hypnotic responding and self-deception

Irving Kirsch

Department of Psychology, University of Connecticut, Storrs, CT 06269-1020. **irvingk@uconnvm.uconn.edu**

Abstract: As understood by neodissociation and sociocognitive theorists, hypnotic responses are instances of self-deception. Neodissociation theory matches the strict definition of Sackeim and Gur (1978) and sociocognitive theory matches Mele's looser definition. Recent data indicate that many hypnotized individuals deceive themselves into holding conflicting beliefs without dissociating, but others convince themselves that the suggested state of affairs is true without simultaneously holding a contrary belief.

The hypnotized person's false belief in a suggested state of affairs is widely seen as a central feature of hypnosis (McConkey 1991). The nature of this delusion has been a topic of intense debate. Sociocognitive theorists (Gorassini, in press; Sarbin 1989; Spanos 1986) view it as an instance of self-deception in much the same way that the phenomenon is conceived by Mele. They propose that hypnotized people convince themselves that the suggested state of affairs is true, but they do not contend that the person simultaneously believes the suggested state of affairs to be false.

In contrast, Hilgard's (1986) rival neodissociation theory fits Sackeim and Gur's (1978) stricter definition of self-deception. According to Hilgard, the hypnotized person holds two contradictory beliefs simultaneously, without being aware of holding one of those beliefs. This is made possible in Hilgard's theory by a division of consciousness into two parts that are separated from each other by an amnesic-like barrier. One part of consciousness intentionally initiates suggested movements, inhibits prohibited movements, and retains full awareness of the actual state of affairs. The other part of consciousness experiences suggested movements as occurring on their own, finds it impossible to make prohibited movements, and is unaware of prohibited memories or sensations. Studies purporting to breach the amnestic barrier, thus allowing contact with a so-called hidden observer, were cited as evidence of dissociation. However, later studies indicated that this "hidden observer" is an experimental creation, rather than an indication of a preexisting division of consciousness (Spanos 1986).

A third possibility is suggested by data reported by Comey and Kirsch (1995). These data indicate that many hypnotized people persuade themselves into simultaneously holding conflicting beliefs, but they do so without dissociating or segmenting consciousness. Rather than being unaware of holding both beliefs, what they seem to be unaware of is the discrepancy between the two beliefs. For example, of 134 people who displayed an apparent inability to bend an arm following a suggestion for arm rigidity, 62 (46%) indicated that they had tried to bend the arm and also that they could have bent their arm if they had really wanted to. Similarly, of 70 people who displayed suggested amnesia, 31 (44%) claimed they wrote down every suggestion they could remember and also that they could have remembered the suggestions if they really wanted to. This might be interpreted as an indication of the tolerance of logical incongruity that has been termed *trance logic* (Orne 1959), except that these ratings were obtained after hypnosis had been terminated. Thus, the self-deception was maintained outside of the hypnotic context.

Although the Comey and Kirsch data indicate that many people acquire contradictory beliefs while responding to suggestions, they also indicate that many do not. Of the participants who responded to the arm rigidity suggestion, 34% maintained that they could not have bent their arm, even if they really wanted to, and 45% of those who displayed amnesia claimed even if they wanted to, that they could not have remembered. Does this indicate self-deception in the weaker sense described by Mele? I think not. Although it is likely that these people, like those who reported contradictory beliefs, have persuaded themselves into believing the suggested state of affairs, there may be no deception involved. Beliefs about one's subjective state can produce that state (Kirsch 1985), a phenomenon that is well documented in the literature on placebos (Kirsch, in press). Thus, believing can make the suggested state of affairs true. These participants could not bend their arms, as long as they believed that they could not bend them, and they could not remember while believing they could not remember.

Acknowledging that instances of strictly defined cases of self-deception might be found, Mele hypothesized that these would be *exceptional* instances, rather than the norm. At first glance, hypnosis might seem like an exceptional situation. However, there are reasons for considering it not to be so very exceptional. First, there is the ease with which some hypnotic suggestions are experienced. The arm rigidity suggestion discussed above, for example, is considered to be a moderately difficult suggestion. However, the 134 people who displayed this suggested response constituted 52% of a volunteer, college student sample. Second, it has been

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

well established that all hypnotic responses can be experienced without hypnosis as well and that the induction of hypnosis merely increases the likelihood of responding (Hilgard 1965). As a result, the hypothesis that suggested responses are due to an altered state of consciousness has been rejected by most researchers in the field (see Kirsch & Lynn 1995).

The many faces of self-deception

Dennis Krebs, J'Anne Ward, and Tim Racine

Simon Fraser University, Burnaby, British Columbia, V5A-1S6, Canada.
krebs@sfu.ca; jward@arts.sfu.ca; tpracine@sfu.ca

Abstract: Those who invoke the word self-deception to represent one phenomenon often argue that those who use it to represent another are misusing the construct. Better to recognize that self-deception is a fuzzy concept that may be used to represent a variety of mental processes and states, and to direct our energy toward distinguishing empirically among its forms and functions.

Self-deception may come in many forms. It may come in weak forms based on ignorance, or in strong forms involving beliefs rigidly retained in the face of incontrovertible contradictory evidence. It may come in cold forms originating from unmotivated aspects of information processing, or in hot forms originating from affectively charged motivational biases. It may involve implicit, unconscious processes, or explicit, conscious processes. And, of particular importance in this context, it may come in forms that involve simultaneously harboring contradictory beliefs. Our efforts should be directed toward distinguishing empirically among the many possible forms of self-deception, not arguing on conceptual bases that some qualify as self-deception, and others do not, or that some forms could not exist.

The classic question about self-deception concerns its existence: Is it "real," or just an interesting possibility? The answer depends on the form of self-deception under consideration. Somewhat ironically, the easier a form of self-deception is to explain in terms of established psychological processes and current models of the mind, the less interesting it is. The reason the contradictory belief form of self-deception has captured so much interest is that evidence for its occurrence implies radical changes in current models of the mind. If we were capable of such intrapsychic inconsistency, our minds could not be structured in the ways in which we are wont to assume.

It is appropriate to adopt a conservative strategy in evaluating evidence for contradictory belief forms of self-deception; however, we must be careful to avoid two mistakes. First, demonstrating that an incident of self-deception could have been determined by a process that does not involve contradictory beliefs does not establish that it did not involve contradictory beliefs. Second, establishing that one incident or form of self-deception does not involve contradictory beliefs does not establish that there are no incidents or forms that do. It follows that we disagree with Mele that the contradictory belief model produces a "fundamentally mistaken view of the dynamics of self-deception" (Abstract), but agree that this form of self-deception is only one of many theoretically possible forms. We also agree that, to date, no one has established that it occurs.

It is important to note that Mele did not prove that the people in his examples did not harbor contradictory beliefs. Indeed, the possibility they did looms like a shadow behind the explanations Mele offers. Clearly, Sam entertained the idea (suspected) that his spouse was having an affair. How far were his suspicions from a belief? Mele offers plausible explanations for the incidents of self-deception he considers, and he cites empirical research in support of the existence of the mechanisms he invokes, but he does not prove that the behavior of people in the garden-variety situations he cites was governed by these mechanisms. Astute scholars could question Mele's explanations and advance convincing arguments that other processes could have produced the effects in question. Without empirical tests of exemplary incidents – ways of establishing what, in fact, is happening in people's minds – we are susceptible to the interminable dances that have plagued the investigation of self-deception for decades.

We believe the resolution of the classic paradox of self-deception will stem from psychological and neurological evidence about the nature of the processes alleged to mediate its various forms. We need to determine what, exactly, a belief is; how people form beliefs, and where, and in what forms, they exist in the brain. We need to develop better models of the self – the knower, or information processor – alleged to be the agent and object of deception. We need to learn more about how information is stored, differentiated, and integrated in the brain and the extent to which independent neural structures process information in parallel ways.

We have a tendency to conceptualize constructs such as beliefs and the self in unitary, all-or-nothing ways: Sam – who has one and only one self – either completely believes his wife is having an affair, or he (all of him) does not. But there is considerable evidence that the human brain is structured in ways that enable people to process different types of (potentially contradictory) information simultaneously. Conscious knowledge is only a small aspect of mental activity. The brain does more than believe – the mental event featured in most models of self-deception – it feels and senses as well. Insight into some forms of self-deception may come with a better understanding of other, perhaps more image-based or affectively charged forms of knowledge, such as those commonly attributed to the right hemisphere. Certainly, split-brain research supplies many examples of people knowing things they deny knowing, though, of course, the patients in question have had their brains cut in half. Déjà vu, jamais vu, and false recognition also appear to involve simultaneously held contradictory beliefs. We sense (believe?) we have experienced events in the past while at the same time believing we have not. People with dissociative disorders involving amnesia, fugue, and multiple identities also appear to compartmentalize information in ways that enable them to believe and disbelieve at the same time.

It is important to ask, "why study self-deception?" One answer is: to enable us to understand better the human mind and human behavior. Another is to help people. The evidence suggests some forms of self-deception are adaptive; others are maladaptive. Health care professionals need to be attentive to the personal and interpersonal functions served by various forms of self-deception. There is good evidence that some forms of self-deceptive optimism, self-efficacy, and idealization foster physical health, psychological well-being, and good interpersonal relations, but there also is good evidence other forms may give rise to pathological conditions like delusions, hallucinations, and dissociative and conversion disorders.

To summarize, Mele makes a good case for the possibility that many garden variety incidents of self-deception may not involve contradictory beliefs, but he does not prove they do not, nor does he establish that people cannot harbor contradictory beliefs. We need to examine empirically the forms and functions of self-deception. As put by Walt Whitman: "Do I contradict myself? Yes, I contradict myself. I am large; I contain multitudes."

Self-deception and the desire to believe

Ariela Lazar

Department of Philosophy, Stanford University, Stanford, CA 94305-2155.
lazar@csli.stanford.edu

Abstract: This commentary concentrates on two flaws in Mele's account. The first is Mele's attempt to account for self-deception by appealing to a desire to believe, together with an instrumental belief concerning the means of satisfying this desire. Contrary to Mele, it is argued that such an account requires a recognition on the part of agents that their actions

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

instantiate these means. Second, Mele misidentifies the most essential – and flawed – ingredient of the standard approach to self-deception, the agent's desire to form the belief (the belief that is undermined by the evidence). This ingredient is retained in Mele's own account of self-deception.

In Mele's example, Don believes that his paper was wrongly rejected by the referees of a professional journal. Such examples are commonly interpreted in the philosophical literature as involving an action that is performed for a reason or with a special kind of intention. Among the philosophers who adhere to this view are: D. Davidson (1986); S. Gardner (1993), D. Pears (1984), A. Oksenberg Rorty (1988), J. Talbot (1995), W. Whisner (1993). This standard approach is popular for the following reason. We are led to understand that Don's aversion to failure in a professional context is high and that a belief that he failed is bound to cause pain and anxiety. To avoid holding this belief or to discontinue holding it (and thus avoid further pain and anxiety), Don attempts to bring about the belief that his paper was unduly rejected. This is an elegant answer to the puzzle of how the belief that is acquired in self-deception is formed and maintained. It accounts well for the following features of self-deception. The self-deceived subject is typically highly irrational – he forms a belief (p) that does not correspond to the evidence at his disposal even when the evidence is overwhelming in support of its negation ($\sim p$). At the same time, the presence of the irrational belief that is formed in self-deception often corresponds to a goal (or goals) of the subject's (e.g., avoiding anxiety, boosting one's self-confidence) while the rational belief conflicts with their satisfaction. As a consequence, many philosophers claim that *the irrational belief is acquired in order to attain a goal that is frustrated by the presence of the rational belief.* Because, by assumption, self-deceived subjects are competent to detect the irrationality of their beliefs, it may seem that this is one of very few available explanations for their formation. The view that the irrational belief is formed with the intention or for the reason of attaining some non-truth-oriented goal presents the formation of this belief as a *consequence of practical reasoning:* it is an outcome of a project that is undertaken by a person to fulfill a desire.

Given the popularity of this approach and its obvious advantages, the question of whether the *belief in self-deception is formed by the agent for the reason of wanting to form that belief* is crucial. Mele ends up offering an account that answers this question in the affirmative in most cases. Although he rejects the view that the self-deceptive belief is formed because of an intention to form it, Mele suggests that, in many cases, it is the desire to believe that, together with some instrumental belief, accounts for the formation of the irrational belief (sect. 4, para. 8 and 10). Rather than focus on the question concerning the relevance of the desire to believe, Mele attacks two features of the standard approach that are mostly by-products of accepting this line of explanation. (In this text I mention only one feature. The second feature namely, whether or not the agent must hold the rational as well as the irrational belief, is discussed in Lazar, forthcoming.) Thus, Mele insists that self-deceived agents in Tversky's experiment for example, do not believe (consciously or otherwise) that they are attempting to shift their tolerance to cold water. This, after claiming that "1. sincere deniers . . . were motivated to believe that they had a healthy heart; 2. that this motivation (in conjunction with a belief that an upward/downward shift in tolerance would constitute evidence for the favored proposition) led them to try to shift their tolerance" (sect. 4, para. 10).

The reader is left puzzled as to how this explanation is completed in Mele's mind when he denies that the agents ever recognize (consciously or not) the nature of their actions: if agents are said to be shifting their tolerance *for the reason* of desiring to hold the belief, how can it be true, at the same time, that they never recognize their actions as instances of shifting tolerance? This is not to be confused, as Mele seems to do, with agents' beliefs concerning *the causes* of their actions. Reason explanations apply easily without the latter belief but not without the former. Take, for example, my refusing to help out a friend in need. I may have no beliefs (or have false beliefs) concerning the actual cause(s) of my refusal. Thus, I may wonder whether my refusal was caused by my promoting my own selfish interests or by the sincere conviction that my friend would be better off in the long run if he were forced to bear the consequences of his deeds. It seems that I recognize my action as both having the features of corresponding to my own selfish interests and to my thoughts concerning the well-being of my friend, but I do not know which set of reasons (if not both) were causally effective in shaping my response. This is quite different from claiming that, given that the operative reason is the promotion of my selfish interests, at no point do I recognize (on any level) that this action is an instance thereof. Pace Mele, this is a conceptual and not an empirical issue.

But it is the idea that the agent's desire to believe is an operative goal in self-deception that constitutes the main flaw in Mele's attempt. Indeed, this is the one essential ingredient of the standard approach and is not disowned by Mele. The idea fails for a number of independent reasons that are discussed fully in Lazar (forthcoming). In this context, I shall briefly make a few comments. First, self-deception is often driven by a strong desire (e.g., to lead a long and healthy life) but ends up being explained by appeal to the desire *to believe* that such a life is forthcoming. In many cases, however, these goals are in conflict. So, whereas the standard approach portrays self-deceived agents as striving to fulfill the goal to believe (that they will lead long lives), it often portrays them as doing that at the expense of satisfying the original desire (to lead long lives). But this does not make sense: after all, the presence of the desire to believe is explained by the intensity of the original desire (e.g., to lead long lives). In this case, an awareness of one's predicament may be instrumental in making it the case that one does lead a long life (e.g., by taking preventive measures).

Other problems with this approach (e.g., its total inability to treat cases of "negative" motivated irrational belief formation such as irrational jealousy or underrating of one's accomplishments) make it a weak contender for accounting for self-deception. Mele's account does not reject the centrality of one's desire to believe vis-à-vis self-deception. In so doing, he maintains the essential ingredient of the standard approach. But it is not the desire to believe, for example, that one will lead a healthy life or that one is physically attractive that drives self-deception but rather the desire to lead such a life. The account of motivated irrational belief formation must be formed around this element.

Distal versus proximal mechanisms of "real" self-deception

Joan S. Lockard

Department of Psychology, University of Washington, Seattle, WA 98195.
jsl@u.washington.edu

Abstract: There is little fear that the concept of *motivational bias* as proposed by Mele is likely to dampen the current academic ferment (see Mele's Introduction) with respect to self-deception for several reasons: (a) like philosophy, science has more recently abandoned the heuristic of a rational human mind; (b) the concept is parsimonious, applicable to many research topics other than self-deception, and, therefore, scientifically serviceable; (c) as a proximal mechanism it addresses process rather than function, that is, *how* rather than *why* questions; (d) it is not as interesting a question as why there is a high prevalence of "real" self-deception (i.e., "garden-variety self-deception" as described by Mele, see sect. 6); and (e) a more penetrating issue is whether "real" self-deception is adaptive.

It is evident that the concept of self-deception (irrational thought, etc.) has gone full cycle in some 30 years, namely, receiving impetus in philosophy in the 1960s (e.g., Fingarette 1969) and resurfacing as a prominent philosophical issue in the 1990s (e.g., Mele) with a more focused revaluation of its definitions and likely

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Table 1 (Lockard).[1]

Examples of proximal processes	Extent of central input	
Sensory:		
Habituation	+	(low)
Selective Perception	++	
Subliminal Perception	+++	
Misperception	++++	(high)
Memory:		
Working memory	+	
Encoding	++	
Long-term storage	+++	
Retrieval	++++	
Restorage	+++++	
Cognitive:		
Emoting	+	
Self-serving	++	
Asserting	+++	
Negotiating	++++	

[1]The problem of understanding proximal mechanisms of "real" self-deception, then, becomes one of determining which multifaceted interaction of sensory, memory, and cognitive levels the individual is executing. Add to this the brain's capacity for lateralized function (e.g., Ojemann 1979) and compartmentalization (brain modules, e.g., Barlow 1989), and a myriad of possible neuronal processes become involved.
Adapted from Lockard & Mateer 1988.

proximal mechanisms. However, Mele's four examples of ways "real" self-deception (see sect. 6) could occur through motivational bias (sects. 3.2.1–3.2.4) are a small fraction of the many possible ways detailed by Lockard and Mateer (1988). As outlined in Table 1, "real" self-deception could involve any of a number of combinations of at least four sensory processes (habituation, selective perception, subliminal perception, and misperception), in conjunction with at least five memory processes (working memory, encoding, long-term storage, retrieval, and restorage), and during any one of several (say, four, for purposes of illustration) cognitive conditions of varying intensity (emoting, self-serving, asserting, and negotiating). Further, the asymmetry of function of the cerebral hemispheres has been well documented with respect to such diverse phenomena as language, spatial organization, handedness, emotion, and cognition (e.g., see review, Springer & Deutsch 1985). Also, lateralization of certain functions such as language manifest gender differences as well (Kimura 1987; Mateer et al. 1982; McGlone 1980; Ojemann 1979). Therefore, the number of possible sensory and memory states, cognitive conditions, and hemispheric asymmetries that could be involved in the neurological processing of "real" self-deception may be 10-fold more than would be hypothesized by Mele's four examples.

Now that we have some appreciation of the complexity with which the normal brain could be engaged in self-deception on a daily basis, let us turn from questions of *how* to more compelling questions of *why*. Why is "real" self-deception so prevalent and seemingly adaptive? If we accept motivational bias as an important proximal mechanism, is Mele's supposition then correct (see his Introduction) that the scientific excitement regarding self-deception becomes greatly diminished? Such a conclusion is reminiscent of the 1950s when the utility of the concept of motivation itself was being questioned by physiological psychologists and, in its stead, a multitude of operationally defined proximal mechanisms was being substituted (e.g., see discussion by Zeigler 1964). Is this yet another lesson we will be forced to repeat? The importance of the concept of self-deception does not rest solely, or even predominantly, on a premature understanding of one category of possible proximal mechanisms. It is a theoretical understanding of the functions that the manifestations of self-deception may serve (Lockard 1980) and the origins from which it may have evolved (Lockard 1978; 1988) that sparks scientific interest. In spite of Mele's model of motivational bias, the science of seeking distal mechanisms of self-deception is well and thriving.

For example, do not the phenomena of self-knowledge, hope, worry, fear, and anxiety suggest that through human brain evolution it has become adaptive to treat oneself cognitively as one would a conspecific (i.e., another member of our species)? There is no question that the ability to argue with oneself, to learn from one's own past experiences, and to plan for a less risky future are adaptive. Then why is it so strange to think that deceiving oneself on occasion as we would deceive another is somehow no more glorious than motivational bias? We would expect adaptive behaviors (including cognitions) to evolve to increase the predilection by which they are learned, the efficiency with which they are executed, and the specificity of context in which they are beneficial; self-deception is not likely to be an exception.

The extent to which the human brain is modularized and lateralized as in the case of language, emotion, and cognition (e.g., Barlow 1989) could, and most likely does, facilitate the proximal processes of self-deception. That motivational bias is operative in self-deception is reasonable and subject to empirical verification, but the distal functions such bias serves, and the extent to which the compartmentalized human brain facilitates the perceptual and neurological processes by which the bias is evinced, are the stuff of which scientific intrigue is made.

In pursuit of distal mechanisms of self-deception, one is reminded of the words of Ekman (1985) that deceit is a way of life and if false information were never conveyed or, alternatively, the truth never told, our emotional lives would be impoverished and more guarded than they are:

> And if we could never lie, if a smile was reliable, never absent when pleasure was felt, and never present without pleasure, life would be rougher than it is, many relationships would be harder to maintain. Politeness, attempts to smooth matters over, to conceal feelings one wished one didn't feel – all that would be gone. There would be no way not to be known, no opportunity to sulk or lick one's wounds except alone. (Ekman 1985, p. 283)

Being a social species demands a middle ground, and through self-deception we can escape the problems and guilt of interpersonal deceit. Surely, "real" self-deception is adaptive and the brain processes that fosters it are likely to have been subject to natural selection. To pursue its theoretical origins and distal mechanisms in comparative species (e.g., Lockard 1978; 1980; 1988; Trivers 1985) and through more astute neurological, cognitive, and behavioral human research (e.g., Sackeim & Gur 1979; Lockard & Paulhus 1988) will undoubtedly compensate for any mundane feelings that may arise from a greater understanding of a likely proximal mechanism such as motivational bias.

Self-deceivers' intentions and possessions

Michael Losonsky

Department of Philosophy, Colorado State University, Ft. Collins, CO 80523.
losonsky@lamar.colostate.edu; www.colostate.edu/depts/philosophy/losonsky

Abstract: Although Mele's four sufficient conditions for self-deception are on track insofar as they avoid the requirement that self-deception involves contradictory beliefs, they are too weak, because they are broad enough to include cases of bias or prejudice that are not typical cases of self-deception. I discuss what distinguishes self-deception from other forms of bias.

Self-deception and bias. It seems that Mele's four jointly sufficient conditions for self-deception do not distinguish between bias

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

and self-deception, or what is typically identified as self-deception. Although self-deception seems to be a species of bias, not all cases of bias are cases of self-deception.

For example, his list of cases in which a desire that *p* be true contributes to believing that *p* is true are mostly cases in which people have a bias or prejudice that brings about the fact that they have a false belief, but they do not look like typical cases of self-deception. Historians who are selective about the evidence they use to support their theses because they want to believe their theses are clear cases of prejudice or bias, but it is not obvious that such historians are also deceiving themselves. Suppose the historians are racists or anti-Semites. Although that belief that Jews or people with darker pigment are inferior satisfies all four of Mele's jointly sufficient conditions for self-deception, their racism or anti-Semitism seems different from the case of parents who refuse to believe that their child is a drug-addict or guilty of a crime.

Biased believers who are not self-deceivers have evidence that provides a greater warrant for the denial of the proposition they believe than for the affirmation of the proposition that they believe, but this evidence need not play much of a role in the believers cognitive architecture. The evidence is there, but simply rejected or ignored on the basis of motivated cognitive mechanisms. Self-deceivers, on the other hand, have this evidence in some stronger way, and this is what theorists such as Davidson (1985) or Sackeim and Gur (1987; 1985; Gur & Sackeim 1979) are trying to capture with the clause that self-deception involves holding two contradictory beliefs. Although I agree with Mele that this condition is too strong, it also seems that Mele's conditions, especially the fourth, are too weak.

How self-deceivers possess evidence. Self-deceivers not only possess the evidence, but it continues to play a role in their belief formation mechanisms. This evidence must at least be such that under ordinary circumstances the evidence would lead to belief in such individuals, but in these extraordinary circumstances this function is blocked or repressed by the believers' motivations. Moreover, the evidence is such that even if it does not trigger belief, it keeps trying to trigger belief and perhaps leads to various proto-beliefs to which self-deceivers have access when probed under the right conditions.

For example, self-deceivers who come to recognize that they were self-deceivers often report that although they had rejected the belief, say, that they were anorexic (or that their children were using drugs), "on another level I knew I was anorexic," or, "I knew all along she was abusing drugs, but I refused to accept it." This sort of admission of denial must be captured in a more accurate account of typical cases of self-deception. Self-deception involves some kind of recognition of the fact that the available evidence warrants the undesirable proposition more than the desirable one. This can be manifested in various ways. One way is in a recurring or nagging doubt that typically does not occur when subjects fix their beliefs. Similarly, self-deceivers can find themselves repeatedly and obsessively entertaining the undesirable proposition and going over the same line of reasoning that supports the desirable proposition. It is as if the cognitive mechanism cannot help but respond to the force of the possessed evidence, although the motivational structure is able to override it.

The self of real self-deceivers. Perhaps Mele is blinded to these sorts of conflicted internal states because he has a notion of a very stable and unified self. Mele considers self-deception in the case of multiple personality disorder (MPD) and rightly maintains that this is not relevant to the understanding of self-deception in more typical cases. However, MPD may not be wholly irrelevant to understanding the healthy self. Perhaps the self is better understood as an organization of various competing and cooperating modules without a central, coordinating processor (Dennett 1991; Minsky 1985). In the normal case, there is sufficient coherence to produce what we consider to be an integrated or mostly integrated self, whereas in the extreme cases of MPD the various modules cohere in ways that mimic very distinct personalities. But even in the normal case, there is sufficient diversity in one's cognitive and motivational structure to allow for various aspects of who we are – for example, who we are at work *versus* who we are at home – and these various aspects can very well be in conflict, particularly when an individual is under some pressure, as is the case when one is facing anorexia or drug-abuse in oneself or in a close friend or family member.

For example, in circumstances where our better cognitive mechanisms can function without too much intrusion – when we have the time and security to be reflective and objective – we can reach an undesirable conclusion, for example, that someone close to us is anorexic. But the moment we find ourselves in day-to-day interactions with this person, we find ourselves denying the obvious and believing what we want to believe. This is not a case of believing a contradiction; rather, our beliefs fluctuate depending on the circumstances we are in. I suggest that this conflicted mental life is an important key to understanding the structure of real self-deception.

Self-deceiving intentions

Mike W. Martin

Department of Philosophy, Chapman University, Orange, CA 92666

Abstract: Contrary to Mele's suggestion, not all garden-variety self-deception reduces to bias-generated false beliefs (usually held contrary to the evidence). Many cases center around self-deceiving intentions to avoid painful topics, escape unpleasant truths, seek comfortable attitudes, and evade self-acknowledgment. These intentions do not imply paradoxical projects or contradictory belief states.

I agree that garden-variety cases of self-deception do not involve a deliberate (self-knowing and fully self-aware) intention to deceive oneself; nor do they involve full-blown contradictory beliefs. Self-deception is not a "reflective project" in which persons self-consciously attempt to convince themselves of what they believe is false (Sartre 1966, p. 89). Contrary to Mele's suggestion, however, garden-variety self-deception is not reducible to nonintentional, bias-generated false beliefs (whether or not held contrary to the evidence; Mele). Despite his nuanced and richly insightful discussion, Mele's "deflationary" position ultimately removes paradox at the cost of being reductionistic and eclipsing intentional self-deception.

Intentional self-deception is ruled out, or relegated to a few rare oddities (sect. 6.3), if we stipulate that the only thing that counts as an *intention to deceive oneself* is the conscious intention to deceive oneself into believing what one also believes is false. But that stipulation is unwarranted. Familiar examples of self-deceiving intentions include: to avoid painful topics, evade unpleasant truths, seek comfortable beliefs and attitudes, and (more generically), disavow or evade fully acknowledging something to oneself (Fingarette 1969; Martin 1986). These intentions can be understood without paradox. To that end, I offer five comments.

First, self-deceiving intentions are carried out by using other intentions. By Mele's own account, self-deception often involves intentional activities such as selective attending and ignoring, selective evidence-gathering, and misinterpreting evidence. Mele omits these activities when he lists sufficient conditions for self-deception, treating them as secondary matters explained by the same motivational biases that cause self-deceptive beliefs. Granted, sometimes these activities are directed entirely by nonintentional biases. Other times, however, the activities function as tactics guided and unified by additional self-deceiving intentions. Occam's razor rightfully prohibits ascriptions of self-deceiving intentions where mere bias is adequate to explain the facts, but much self-deception can only be understood as the intentional evasion of unpleasant topics and truths.

Second, self-deceiving intentions are, for the most part, (purposefully) kept nonreflective. They remain either "prereflective" in Sartre's sense (conscious but not self-conscious), not "spelled

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

out" in Fingarette's sense, unconscious in Freud's sense (dynamically repressed), or unconscious in Audi's sense (extremely difficult to bring to consciousness, Audi 1985, p. 174; Fingarette 1969). Only in retrospect or in emerging from self-deception will individuals interpret their complex patterns of behavior, emotion, and reasoning as products of self-deceiving intentions, although as observers we can sometimes make this interpretation earlier.

Third, Mele is right that self-deceiving intentions do not require ascribing full-blown contradictory beliefs. He fails to allow, however, that a suspicion, partial belief, fear, or hope (often nonreflective) can suffice as the epistemological basis for forming and carrying out self-deceptive intentions (Butler 1896). Familiar descriptions of self-deceivers in folk psychology testify to something less than contradictory conscious beliefs being present: "She suspected (partly believed, knew) deep down, in her heart, that her husband was having an affair, but refused to look honestly at the evidence." Nor do self-deceiving projects require a metabelief that one has strong suspicions contrary to what one wants to believe. At most there might be a momentary (and then purposefully ignored) sense that one is not being completely honest with oneself.

Fourth, although self-deception does not involve reflective projects, we are not misguided in thinking of self-deception as an intrapersonal analog of interpersonal deception. Even interpersonal deception does not require a reflective intention ("I am engaging in deception"). Because "deceive" may carry pejorative connotations, or at least raise questions about prima facie wrongdoing, interpersonal deceivers frequently accent other intentions involved. For example, the mother, spouse, or friend who deliberately misleads someone they care about may construe their intentions simply as being to protect, support, or help someone they care for. Here there is a parallel with self-deceivers who act with purposes, strategies, and intentions they do not acknowledge (to themselves or to others) as intentions to deceive.

Numerous additional analogies are worth exploring. For example, although self-deception does not involve fully conscious contradictory beliefs, typically it does involve a cognitive conflict, for example, suspecting *p* and believing not-*p*. Self-deception need not be exactly like interpersonal deception, any more than teaching oneself is exactly like teaching others, to justify exploring such analogies (Gardner 1969–70, p. 243). (I speak of exploring analogies, not of strictly modelling self-deception on interpersonal deception.) In any case, there remains a striking analogy in how both self-deceivers and interpersonal deceivers evade acknowledging truths (to themselves or to others).

Finally, scholarly vignettes of self-deceivers can be like drawings of duck-rabbits: the details are sufficiently sketchy to justify alternative interpretations, and one's interests can influence whether one sees a bird or mammal. Thus, the cuckolded husband who believes his wife is not having an affair can be interpreted as a nonintentional or an intentional self-deceiver, depending on how the details are fleshed out. "Real" cases are also open to alternative interpretations, especially where both intentional evasion and nonintentional bias are present in the same case. We do best to begin with elaborately described examples from clinical studies and from literature, for example Casaubon in Eliot's *Middlemarch,* Karenin in Tolstoy's *Anna Karenina,* and Judge Pyncheon in Hawthorne's *The House of Seven Gables.* All this complicates scientific studies of even garden-variety cases, but then self-deception is complicated, as the enormous literature devoted to it testifies. Much, too, depends on whether our garden is overrun with ducks, rife with rabbits, or home to the duckbill playtypus.

Direct, fully intentional self-deception is also real

Christian Perring

Department of Philosophy, University of Kentucky, Lexington, KY 40506-0027. **cperring@ukcc.uky.edu**

Abstract: An important way to become self-deceived, omitted by Mele, is by intentionally ignoring and avoiding the contemplation of evidence one has for an upsetting conclusion, knowing full well that one is giving priority to one's present peace of mind over the search for truth. Such intentional self-deception may be especially hard to observe scientifically.

Mele considers two sorts of intentional self-deception. The first is given in the case of Ike, who decides to fool his future self by planting false evidence. This is an indirect method of changing one's beliefs. Mele says this form of belief-manipulation is rare, and he is surely right about this. The second sort of intentional self-deception considered is via "hidden" intentions, not fully conscious or epistemically accessible to the person. Mele says that this model is theoretically perplexing and is a problematic conception of self-deception, he has a simpler model that is about to account for the phenomena. Mele leaves us to draw the conclusion that, using good scientific method, we should adopt the simpler model, other things being equal.

Consider the case of Sam and Sally that Mele gives us in section 3. Sam uses negative misinterpretation, positive misinterpretation, selective focusing, and selective evidence gathering to maintain his belief that Sally is not having an affair. Mele argues that these forms of biasing the evidence do not have to be intentional. Indeed, he points out that if one knows that one is biasing evidence, then it will be very hard to believe the results of one's deliberations. Mele says we do not have to suppose that Sam is intentionally protecting his favored belief that Sally remains faithful to him to understand his self-deception. I aim to show that this is wrong, and that there is a form of intentional self-deception he has not considered.

Let me add a little to the sorry tale of Sally and Sam. Sam has been married and divorced twice before. Both divorces were bitter. He is very busy in his work, and he is exhausted when he gets home. When his close friend tells him about seeing Sally with Mr. Jones, he shudders in fear and anxiety in recognition of the potential significance of the information. He has strong religious beliefs and could not stay with Sally if he knew she were unfaithful. He says to himself, "I am not going to think about this." He turns the TV on, drinks a few beers, and does not think about the evidence again that evening, or indeed, for several weeks, until he is directly faced with it again.

I have tried to describe a familiar case of a person who is unwilling to face the facts. Avoidance and denial are common forms of self-deception. Sam does this to maintain his calmness and to avoid the pain of thinking about another divorce. His intentional self-deception is not self-defeating, but this is not because the intention to avoid thinking further about the evidence he has is hidden from him. It is quite explicit, and at the time he makes the decision to avoid assessing the evidence he has, the search for truth takes second place to his need to maintain psychological equilibrium. The self-deception is successful because he does not reflect on it afterward, but rather immerses himself in other activities. It is the initial avoidance and the initial engagement that constitute the intentional self-deception. He remains self-deceived for several weeks through the luck of not being faced with the crucial evidence again, and possibly through unintentional processes that maintain his false belief. The intentional self-deception here is not paradoxical or theoretically perplexing.[1]

Consider another simpler case. Jim, who is 12 years old, starts telling (rather maliciously, but truthfully) his 8-year-old brother Andrew that his pet rabbit has died. Andrew puts his hands over his ears, yells out loud, and runs from the room. Andrew was worried that there might be bad news, and so he blocked it out,

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

and maintains his belief that his rabbit is alive. Andrew loved his rabbit, and clearly is engaging in his avoidance of the bad news because he knows that bad news is in the offing. This is selective attending to evidence of the most intentional kind. It does not always lead to successful self-deception, but our experience tells us that it often does. Intentional self-deceivers manage to maintain the belief that not-*p* by intentionally ignoring, blocking out, and engaging in activities that will lead them to forget evidence that *p*. Note that this self-deception is not just a form of sticking one's head in the sand and avoiding all information. Both Sam and Andrew do have distressing evidence given to them, but they find ways to forget that information. Mele's model of self-deception as unintentional mistakes resulting from our fears and desires does not explain the behavior of Sam or Andrew in the scenarios above, where their actions to avoid evidence for upsetting beliefs are quite deliberate.

Finally, a point about observing self-deception. I conjecture that intentional self-deception occurs mostly when a person gains access to very upsetting evidence. The less emotionally charged experiments that Mele does discuss in his target article are more likely to be explicable just by his motivated mistakes model. If this is right, then there is a specific problem in measuring intentional self-deception. We cannot create such experimental situations in psychology laboratories, with volunteer subjects being given hurtful information, if only because of the ethical restrictions placed on researchers. A more profitable way to research intentional self-deception might be to observe the reactions of people getting bad news in real life, but even then subjects are unlikely to be willing to answer a researcher's questions about what they were thinking when they got the news, and whether they intentionally engaged in denial when first hearing the news. The subjects' reports of their thoughts are likely to be unreliable anyway, especially because they may still be engaging in self-deception. So we probably have to get the best descriptions of intentional self-deception from our everyday experiences of avoidance and denial.

NOTE

1. For more detail, see my Ph.D. dissertation *The Limits of Irrationality* (Princeton University, 1996).

The uses of self-deception

Howard Rachlin[a] and Marvin Frankel[b]

[a]Department of Psychology, State University of New York, Stony Brook, NY 11794. **hrachlin@psych1.psy.sunysb.edu** *[b]Department of Psychology, Sarah Lawrence College, Bronxville, NY 10708*

Abstract: The essence of a mental event such as self-deception lies in its function – its place in the life of an animal. But the function of self-deception corresponds to that of interpersonal deception. Therefore self-deception, contrary to Mele's thesis, is essentially isomorphic with interpersonal deception.

Mele's target article considers three approaches to "the nature and (relatively proximate) etiology of self-deception." But all three approaches are cognitively based. Without a prior account of the *function* of self-deception – the place of self-deception in human (and nonhuman) life – the efforts of Mele and the philosophers and cognitive psychologists he criticizes are analogous to attempts to understand how a chair is made without first understanding that chairs are made for sitting. What function might self-deception have in a person's life?

The main issue addressed by Mele – whether self-deception is isomorphic with interpersonal deception – has meaning only if we first consider the extent to which the functions of self-deception and interpersonal deception overlap.

To some extent, at least, their functions do overlap. Some actors believe that they can convince the audience that they are really feeling an emotion only if (for the duration of the performance) they first convince themselves that they are feeling it. From this viewpoint, self-deception is nothing but a form of imagination – behaving as if a certain situation existed when it does not in fact exist. A common function of such behavior in everyday life is to bring about the very situation imagined. The paradox of self-deception is not that it requires us to believe *p* and not-*p*, but that, when successful in its natural function, self-deception becomes veridical perception. A famous "method" actress, for example, hated Hollywood parties but felt she had to go to them to advance her career. After experiencing several such parties as torture she realized that after all she was an actress and decided that at future parties, come what may, she would act as though she were enjoying herself. Almost immediately after putting this plan into effect, she came to believe that she *was* enjoying herself. Is this self-deception? It would seem to fit Mele's definition: her motives biased her perception. But in this case, a useful function was served and ultimately her perception was veridical. Some of Mele's own examples may be interpreted in this way. Sam, who believes "for years" that his wife Sally would never have an affair, when she actually is having affairs, is a case in point. Sam's belief is unrealistically biased by his hopes. According to Mele, he is therefore self-deceived. But if he has been able to maintain this self-deception *for years*, Sally must have been at least discreet; she must have been carefully fitting her affairs into her marriage. Sam's persistent self-deception may never bring about Sally's fidelity, but it may well maintain a marital state as blissful as if Sally were faithful. If this were the case (not that it must be) Sam's self-deception would have enabled him to overlook a truly irrelevant detail about his relations with Sally and to focus on the important aspects of his marriage. Sam would have essentially imagined his way to a happier life. Without this self-deception Sam, irrationally overcome by jealousy or shame, might have divorced Sally who, in every meaningful way, might have been a perfect wife.

Interpersonal deception often serves a similar function. By not telling Sam what they know about Sally's behavior his friends could be helping him to achieve the same higher good that his self-deception serves. From their point of view, their actions serve a social good – preserving Sam and Sally's marriage and providing for their children. But you do not have to look to overly idealized marital situations to find self-deception functioning beneficially. People who walk for blocks through streets full of litter holding a gum wrapper to deposit in a trash basket (and many do) are essentially imagining a degree of social cooperation that does not exist. They are thereby self-deceived. However, refusal to litter may fit into a wider pattern of virtuous acts that form a coherent and highly functional self-concept. Self-deception in a narrow context (a walk in the street) may be a necessary part of self-understanding in a wider context (the person's whole life).

The purpose of a belief is to guide our actions. Most of the time, when we act against our immediate interests because we believe we will bring about a higher good that does not currently exist, we are, by Mele's criterion, self-deceived. A person may vote in a national election, for instance, because he believes that his vote makes a difference. But one vote virtually never makes a difference. Therefore every person who votes (believing his vote makes a difference) is deceiving himself. But here again self-deception is functioning normally – as it is supposed to function. To preserve *that* function, and not, as Mele implies, because of some quirk in our cognitive make-up, is *why* we deceive ourselves.

Only after we have understood the normal function of self-deception can we look for subversions of that function, as illustrated by the Quattrone and Tversky (1984) study. Just as you might deceive yourself into a belief biased by your long-term good, so you may deceive yourself into a belief biased by your short-term good (and at the expense of your long-term good). The subjects who shifted their pain tolerance levels in accordance with their hopes rather than their fears avoided the trouble of further medical tests and the possible pain and inconvenience of medical treatment. The price was the possibility that their fears were really justified. Had Quattrone and Tversky conducted their study with a

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

group of hypochondriacs as subjects, we would expect an opposite bias in tolerance thresholds. Why? Because hypochondriacs by definition value the benefits of being unhealthy (attention from doctors and family, days off from work, etc.) more than the costs.

Of course the sort of self-deception studied by Quattrone and Tversky has absolutely no chance of bringing about the hoped-for conditions and is therefore self-destructive as well as self-deceptive. This form of self-deception is the one Mele concentrates on but, functionally, this form also has parallels in interpersonal deception. Just as we say deceive ourselves for our own greater good we may deceive others for their own good, or for the good of society. On the other hand, just as we may deceive ourselves for our own immediate good we may deceive others for that same reason.

Mele is right that self-deception is always motivated, but *all of our beliefs are motivated*. We are self-deceived, not when our beliefs are motivated, but when they are contrary to the present state of affairs. Our judgments are *always* influenced by our hopes. Contrary to Mele's thesis, there is no such thing as "cold" biased belief. All four of Mele's examples of such "cold" biases are clearly motivated. We pay more attention to vivid than pallid information, for example, because vivid information is usually more important. There are always reasons why a belief is biased one way or another. A psychologist's job is to find those reasons.

ACKNOWLEDGMENT

This commentary was prepared with the assistance of an NIH grant.

Flavors of self-deception: Ontology and epidemiology

Harold A. Sackeim[a] and Ruben C. Gur[b]

[a]*Department of Biological Psychiatry, New York State Psychiatric Institute and Department of Psychiatry, College of Physicians and Surgeons, Columba University, New York, NY 10032* **has1@columbia.edu;**
[b]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104 **gur@bbl.psycha.upenn.edu**

Abstract: Mele questions the prevalence and ontological status of strong forms of self-deception, as well as our attempt at experimental demonstration. Without validated indicators outside laboratory contexts, statements about prevalence are purely speculative. Conceptualizing self-deception without positing the motivated lack of awareness of a contradictory belief is unsatisfactory in dealing with issues of "agency," that is, how can we stop the processing of threatening information unless we recognize that the information is threatening?

Mele offers a set of conditions that he argues are sufficient for ascribing self-deception and that do not require intentionality or that the self-deceived simultaneously hold contradictory beliefs. Mele claims that his description is adequate to account for everyday instances of self-deception, and that, indeed, if they occur at all, cases of self-deception involving mutually contradictory beliefs are exceptional. In essence, Mele argues that self-deception should be thought of as but one example of how individuals acquire or retain motivationally biased beliefs. As Mele notes, the import of this work is "deflationary," because traditional notions of self-deception typically require that individuals hold contradictory beliefs and that motivational factors determine which belief (p or $\sim p$) is subject to awareness. In rejecting the motivated lack of awareness of beliefs as a necessary component of self-deception, Mele takes a step toward questioning the conceptual necessity of positing a "dynamic unconscious."

Ontology. Mele acknowledges that strong forms of self-deception, as defined, for example, by the criteria we offered as necessary and sufficient (Sackeim & Gur 1978; target article, sect. 4), are conceptually coherent. He claims only that most instances of self-deception may be understood without requiring the mental states implicated by strong forms of self-deception. There have been a variety of attempts like Mele's to back away from strong forms of self-deception. For example, Fingarette (1969) argued that the self-deceiver fails to spell out his engagement in the world – that is, fails to attend to awkward evidence or derive the necessary conclusions. Greenwald (1988) was also disturbed by the notion of individuals simultaneously holding contradictory beliefs and, using an information processing metaphor, suggested that many instances of self-deception involve limiting attention to or early termination (prebelief) of the processing of unpleasant information.

We do not doubt that weak forms of self-deception are conceptually coherent and have empirical reality. However, the strong form of self-deception, based on the analog of interpersonal deception, is specifically of interest because it requires a partitioning of consciousness such that individuals are capable of simultaneously holding contradictory beliefs and, because of motivational factors, are unaware of one of these beliefs. Thus, the strong form of self-deception requires the motivated lack of awareness of the end products of cognition, that is, the establishment and maintenance of beliefs. Regardless of its prevalence, if the strong form of self-deception has any empirical instantiation in normal functioning, there are profound implications for views of consciousness.

Part of the issue in distinguishing between weak and strong forms of self-deception concerns the "agency" involved in aborting the processing of threatening or otherwise counter-motivational information. It is difficult to see how one can effectively "disattend" to threatening information or fail to derive obvious implications from this information unless, at some level, the information is recognized to be threatening. Does the narcissist simply fail to notice signs of failure or, because of the recognition of its threatening nature, does the narcissist deliberate about these signs and reinterpret their significance? Surely, both occur. At issue in understanding the mechanics of weak forms of self-deception is how one knows when and how to stop thinking, without the recognition of where one's thoughts are leading? How can information be experienced as threatening if there is no contradictory belief? If there is recognition of the implications, are we that far from establishing belief (at least at the probabilistic level)? The strong form of self-deception solves the problem of agency by positing that narcissists are not secure of their worth and lack awareness of this belief in personal inadequacy. Weak forms of self-deception mainly differ from strong forms in stopping short of establishing the belief of inadequacy.

Mele challenges the validity of previous attempts to demonstrate the empirical reality of strong forms of self-deception. The heart of his criticism of our early work (Gur & Sackeim 1979; Sackeim 1983) was that physiological or other behavioral indices are not indicators of belief. This position can be interpreted as only admitting self-reports as evidence of beliefs. Because strong forms of self-deception assume that the belief at issue is not subject to awareness, that is, not available to self-report, Mele's position can be viewed as making empirical tests of the ontological status of strong forms of self-deception conceptually impossible. Mele's position is at odds with the decades of research using physiological and other indicators in the detection of interpersonal deception, that is, lie detection. Furthermore, this narrow definition of belief precludes the study of evolutionary mechanisms for self-deception.

Epidemiology. Empirical approaches to the detection of interpersonal deception are probabilistic. We also lack valid methods of detecting self-deception, let alone its flavors, in all but constrained experimental circumstances. Statements about the prevalence of weak and strong forms of self-deception are therefore a matter of prejudice or preference, without empirical basis.

The role of belief. In our view, a key issue in considering the nature of self-deception has not been raised. Beliefs are but one aspect of mental contents and people may not only lie to themselves about to their beliefs, but, perhaps with equal or greater frequency, may be self-deceived about their loves, hates, wishes,

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

and fears. Individuals who deny a heterosexual or homosexual attraction may consciously avow that they are not attracted to X, yet they are. In a strict sense, they do not simultaneously hold contradictory beliefs because the attraction to X is not believed. At issue is the distinction between "John being attracted to X" and "John believing he is attracted to X." Even here, however, the issue of agency is critical. One must ask why a mental content that is usually subject to awareness, such as a sexual attraction, in this case is not. If motivational factors are determinative (unlike instances of blind sight) and the lack of awareness serves to maintain a conception of the self (i.e., a set of beliefs) attribution of the strong form of self-deception may be warranted.

Pragmatically pristine, the dialogical cause of self-deception

Colin T. Schmidt

The Sorbonne University, 75231 Paris, France. **coschmi@idf.ext.jussieu.fr**

Abstract: Empirical evidence of self-deception's propositional duality is not sought; philosophically relevant links between propositions proper and mind are explored instead. Speech in unison ably indicates the social grounding of such attitudinal structures. An extra-theoretical eye – with regard to cognitivism – is cast on a case of "illusory communication." The reinforcing of lexical analysis shows Mele's approach to be in need of non-*ego* concepts, wherefore it lacks soundness with respect to reference.

Self-deception must have *something* irresolvably paradoxical to it, otherwise philosophers would not write dissertations on it. I fully agree with Mele that the dual belief hypothesis *theoretically* implies an impossible contortion of the mental. Hence I have not the least basis for condoning incongruous Intentionality empirically. This said, pragmatics has been neglected. Formulating views on self-deception might be less enigmatic if Mele withdrew his "denotative hooks" from it.

Justifying such a charge could prove difficult. Nevertheless, creating doubt as to the plausibility of blowing out the underpinnings of the interpersonal will suffice because the author sees weakness in modelling self-deception after stereotypical interpersonal deception. A portrayal in which not-so-stereotypical "interhuman" deception constitutes self-deception would directly support the interlocutivity of the phenomenon. Let us set the scene: people oscillate between two viewpoints on computational artefacts, personifying them and downplaying their resemblance to ideal cognitive playmates. (1) When adopting the intentional stance, they project their own propositional attitudes onto them *as if* they were human. (2) When "funny" behaviour (linguistic awkwardness) or dysfunctioning (sparks and flames) occurs reaffirming that they are *just* machines, people adopt the design stance for understanding, or else seek repair (Dennett 1987).

> Inspired by human communication, the computer company Pear Incorporated is devoted to creating perfect mates for lonesome hearts. Herbert believes his brand new Pear Computer™ is so sophisticated that its desires, beliefs, intuition, affectivity, inferential capabilities, and so on will intimately match his own, even though he is unaware of Pear's initiative. However withdrawn Pear's design stage seems from Herbert's present experiential reality, we will assume they [the programmers] mean no evil. Unwrapping it, Herbert exclaims "Bertha! We can now enter the realm of interpersonal communication!" But the anguish and frustration Herbert expresses later as he downwardly revises his model of the depth of Intentionality behind Bertha's interface shows "she" lacks the interpersonal skills he expected. So no cherished "other half" for Herbert (sob . . .).

Peculiarly, it appears Herbert's definition of "the realm of interpersonal communication" seeks to establish their relationship on *information flow*. However, his musing about interpersonal communication was clearly a "communicable" – it is the result of (not the reason for, cause of, or premiss to) Herbert's relational coupling with Bertha, a supposed genuine soul, *for* communication. Had he not deceived himself? Mele makes no mention of the nature of the relationship of the deceived to *p*, which remains simply interpreted as pure propositional content detached from mental states (Engel 1992). But the ambiguity of definitions for "interpersonal communication" involves such mental paraphernalia; most readers will have attributed their own set of propositional attitudes to Herbert – perhaps "wrong beliefs." Other than physically, was Herbert truly alone?

Although lexical positions question the existence of self-deception, they can set the agenda for encompassing the problem of reference in divergent definitions if illustrated in suitable light. Assume the mental acrobatics in question practicable. Whether believing one's own eyes or trusting fellowmen in maintaining the simultaneity of *p* and $\sim p$, consciously or not, the believer is the subject of the verb "to believe"; and likewise for the deceiver, whether deceiving intentionally or otherwise, they are both agents. The link is strong between the pronoun "I," "me," and the Self in what each refers to. I say, "You have deceived me!!!" We all don pronominal labels. Intuitively, early century thought concerning *ego's* privately accessing the referent in communication *with* others holds little water because deceivers are "I"s, too. Mele's *informavorous* system only allows for a "referential" thrust that is singular (Jacques 1990): hence my denotative jab at the outset. If the system were opened to coreferential processes, could he not consider self-deception as isomorphic to its interpersonal counterpart? Whether one labels Herbert's verbal behaviour as the entity ensuing from "inner dialogue" or "holistic belief formation through interacting mental states," or otherwise, dialogically grounding his deceived state with the actions of a putative deceiver would deftly sweep toward reinstating the social dimension self-deception requires. According to French philosopher Francis Jacques (1979; 1985), the Other authorising Herbert's allocution would be of the general sort (not a "peripheral mate"), in essence a structure that organises the perceptual foundations to *Herbert's* window on the world.

Downplaying logico-linguistic analysis makes the entire deceiving process look like a concoction of Herbert's proximal stimuli. What about the distal origin? If we grasp the realities of referential opacity and corporeal experience so crucial to our navigation in life's adventure of avoiding/attracting self-delusion, what Herbert's "organ of thought" would seem to be in the spirit of the target article could escape, as it were, from Putnam's demonstrative cerebral confinement scheme (Putnam 1981).

Individuality stems from the Self being autonomous. But notions themselves are relation-bound; autonomy cannot exist without dependency, just as there is no Self *in absentia* of the Other. Extirpating the Other from the system is at the expense of self-deception's required selfhood. Reexamination of "strong psychological posturing" is necessary (Schmidt, in press); how can any cognitive phenomenon be studied, defined, and so on, according to proper intersubjective criteria if in the throes of mentalism? Herbert's discerning a "mirage-mate" for communicative purposes shows structural intertwining with the interactive reality of interpersonal deception.

My purpose was to illustrate the concomitant nature of reference with which any position on self-deception should endeavour to cope. It would have to exceed the individual rational Self and excel in nontangible social space. Mr. Mele will surely have something to say about my holistic "hocus-pocus" not having said just *how* to activate it. He may have an explanation to the problem exposed, or he may just take it for light cajolery and prompt my further thought on the matter. In any event, I, too, am pessimistic about adequately *documenting* such things as holding opposing beliefs. There are undoubtedly indications of its existence though. Has Mele not managed to hoodwink himself in believing that self-deception safeguards fewer mysteries for rainy days (p) than he knows it does ($\sim p$)?

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Does self-deception involve intentional biasing?

W. J. Talbott

Department of Philosophy, University of Washington, Seattle, WA 98195-3350. **wtalbott@u.washington.edu; weber.u.washington.edu/~wtalbott/**

Abstract: I agree with Mele that self-deception is not intentional *deception;* but I do believe that self-deception involves intentional *biasing,* primarily for two reasons: (1) There is a Bayesian model of self-deception that explains why the biasing is rational. (2) It is implausible that the observed behavior of self-deceivers could be generated by Mele's "blind" mechanisms.

I agree with Mele that self-deception is not intentional *deception.* But I do believe that self-deception involves intentional *biasing,* that is, intentional biasing of one's cognitive processes in favor of the self-deceptive belief that *p.* My comments here must, of necessity, be brief. For a fuller discussion of the issues raised in these comments, I refer the reader to Talbott (1995).

Before I turn to self-deception, let me briefly say something about acting intentionally. I believe that many philosophers have seriously misunderstood the nature of intentional activity, because they have assumed that, as Mele puts it, "doing something intentionally entails doing something knowingly" (sect. 2). Call this the *Transparency of Intentions Thesis.* Mele's discussion of the Quattrone and Tversky (1984) experiment and of his Ann example in section 4, show that he does not accept the Transparency Thesis. Neither do I. So it seems to me that Mele and I are in agreement that doing something intentionally does *not* entail doing it knowingly. But I still have a worry that Mele's notion of intentions as executive attitudes toward plans (sect. 5) would make intentions more premeditated than they need be.

If, as I believe, persons can be ignorant of or mistaken about the beliefs and intentions that are responsible for their intentional activities, then there is no simple test for determining their intentions, or for determining when an action is intentional. As I see it, to say that an action is intentional is to say that it is based on agents' beliefs and desires in a way that is responsive to evidence and reasoning, including reasoning about how to achieve what they desire (none of which need be conscious). The best formal model of intentional action that I know of is found in Bayesian decision theory. In Talbott (1995), I use Bayesian decision theory to explain the sense in which I believe that self-deception is intentional.

In section 3, Mele provides four examples of recognizable self-deception that can easily be explained without supposing that the subject has a desire for a false belief. In each case, the self-deceptive belief that *p* is a proposition that the subject desires to be true. In each case, the subject's preferences rank the possible state of affairs <[I believe that *p*] and [*p* is true]> above the possible state of affairs <[I believe that *p*] and [*p* is false]>. What is characteristic of all four examples, and seems to me to be characteristic of cases of self-deception generally, is that the subject desires to believe that *p, regardless of whether it is true.*[1]

Mele is willing, at least for the sake of argument, to suppose that the desire to believe that *p* regardless of whether it is true does explain some cases of self-deception (note 35). In addition, he agrees that the self-deceptive desire leads to two types of biasing: internal biasing and input-control biasing (sect. 5). But Mele does not believe that the biasing involved in self-deception is typically intentional (note 14). I have two reasons for thinking that the biasing in self-deception *is* largely intentional:

(1) As I show in Talbott (1995), when certain usually uncontroversial auxiliary assumptions are satisfied and subjects desire to believe that *p* regardless of whether it is true, if subjects have a choice between biasing their cognitive processes in favor of *p* or not biasing them, the Expected Utility of the biasing alternative will be greater than the Expected Utility of the nonbiasing alternative. Thus, a Bayesian model would predict that, in such a situation, subjects would at least try to bias their cognitive processes in favor of *p.* It seems to me then that we should expect at least some intentional biasing in such cases.

(2) In Mele's model, there really are no true strategies of self-deception. There are only apparent strategies. For Mele, what seem to be strategies of self-deception are in reality simply the unintended byproducts of "blind" processes triggered by the relevant desires. This seems to be the right kind of account for wishful belief, where there is no tendency to attribute any sort of strategy to the subject. But this sort of account of self-deception strains credulity. For example, as explained more fully in Talbott (1995), one of the most striking elements of the behavior of many subjects with a self-deceptive belief that *p* is the resourcefulness with which they resist the conclusion *r: r* = My belief that *p* is the result of a desire to believe it regardless of whether it is true. I believe that subjects have a reason for this resistance: if they were to accept *r,* this would tend to undermine their belief that *p.* The resourcefulness and ingenuity of self-deceivers in avoiding the belief that *r,* that is, in avoiding coming to the conclusion that their belief that *p* is caused by a desire to believe it regardless of whether it is true, is what most inclines me to think that there are genuine strategies of self-deception. These are not strategies in the sense of consciously premeditated plans, but in the weaker sense of intentional biasing of one's cognitive processes, both internal biasing and input-control biasing (biasing in favor of *p* and in favor of $-r$).

Space constraints prevent me from further elaborating on what seem to me to be the genuine strategies of self-deception, but I expect that most readers will have had some opportunity to observe them, for example, in the variety of techniques of selective quotation, reinterpretation of apparent conflicting evidence, and intimidation that the white supremacist Freemen (or other "true believers") use to "justify" their doctrines. But I acknowledge that we are still a long way from being in a position to design a decisive experimental test of the claim that the biasing is intentional.

NOTE

1. *Desiring to believe that p regardless of whether p is true* is a notion that can be defined precisely in terms of Bayesian theory: consider the subject's preference ranking over the following partition of possible outcomes: (1) <[I believe that *p*] & [*p* is true]>; (2) <[I believe that *p*] & [*p* is not true]>; (3) <[I do not believe that *p*] & [*p* is not true]>; (4) <[I do not believe that *p*] & [*p* is true]>. Someone who desires to believe that *p* regardless of whether *p* is true is someone who ranks (1) and (2) above (3) and (4). By contrast, we can define *desiring to have an accurate belief about p* as having a preference ranking that ranks (1) above (4) and (3) above (2).

Author's Response

Understanding and explaining real self-deception

Alfred R. Mele

Department of Philosophy, Davidson College, Davidson, NC 28036. **almele@davidson.edu**

Abstract: This response addresses seven main issues: (1) alleged evidence that in some instances of self-deception an individual simultaneously possesses "contradictory beliefs"; (2) whether garden-variety self-deception is intentional; (3) whether conditions that I claimed to be conceptually sufficient for self-deception are so; (4) significant similarities and differences between self-deception and interpersonal deception; (5) how instances of self-deception are to be explained, and the roles of motivation in explaining them; (6) differences among various kinds of self-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

deception; (7) whether a proper conception of self-deception implies that definitive ascriptions of self-deception to individuals are impossible.

I am grateful to the commentators for their time and effort. Although I will concentrate on the objections they raised, I also appreciate the support they offered.

R1. My parting challenge: "Dual beliefs." I concluded the target article by challenging critics to provide convincing evidence of the existence of instances of self-deception that involve simultaneously believing that *p* and believing that ∼*p* (the "dual belief" condition). That is an appropriate issue with which to begin here. Explicitly leaving it open that the simultaneous possession of such beliefs is conceptually and psychologically possible (sects. 4 and 6), I argued that influential empirical work on the topic does not meet the challenge and that there is no explanatory need to postulate "dual beliefs" either in familiar cases of self-deception or in the empirical studies discussed. (**Krebs et al.,** who grant that the challenge has not yet been met, mistakenly imply that I attempted to "establish that people cannot harbor contradictory beliefs.")

The thrust of some of the commentaries that directly or indirectly address my parting challenge is that certain empirical or theoretical results provide direct or indirect support for the idea that mental operations are layered, partitioned, or segmented in a way that favors the possibility or probability of someone's believing that *p* while also believing that ∼*p*. I myself would like to see convincing evidence that this dual belief condition is satisfied in some cases of self-deception. Such evidence would settle one significant question about self-deception, and it might even provide indirect support for my own claim that if there is self-deception of the dual belief variety, it is remote from garden-variety instances. However, as I will argue, none of the commentators has provided or cited such evidence.

In preparation for this, some misimpressions should be laid to rest. First, I find nothing in the target article to suggest that I tacitly assume "that people have only one belief about a given topic at a given time" (**Brown & Kenrick**). Given everything I believe now about my children (or self-deception, or the United States), if I could have only one belief about this topic now, its propositional content would be incredibly large! Second, I have always been happy to grant that the large collection of propositions believed by a person at a time may include inconsistencies (**Bermudez, Brown & Kenrick, Foss**). For example, the propositions someone believes now might include a collection of the following sort: if *q* then (*r* or *s*); *q*, *t*, and *u;* if (*t* or *u*) then ∼*r;* if *u* then ∼*s*. My concern is with believing that *p* (e.g., that Bob has had an affair) while also believing that ∼*p* (that Bob has not had an affair), because many have alleged that precisely this condition is necessary for self-deception (see sects. 2 and 4 and Mele 1987b, pp. 2–8). Third, possessing a body of data that provides greater *warrant* for ∼*p* than for *p* should not be confused with believing that ∼*p*. (**Gergen** apparently confuses the two.)

Brown & Kenrick contend that "the *simultaneous* possession of logically contradictory beliefs can . . . be explained without any mysterious cognitive tricks." It is important to be clear about terminology. Some people use the term "belief" to refer both to *what* is believed (e.g., that Bob has had an affair) and to the associated *state of mind* (e.g., Ann's belief that Bob has had an affair). As long as the two senses are not confused with one another, discussion can proceed smoothly. The propositions *p* and ∼*p* are logically contradictory; that is, it is logically impossible that both *p* and ∼*p* are true. This does not entail that it is logically impossible to believe that *p* while also believing that ∼*p*.

Brown & Kenrick offer alleged examples of "logically contradictory beliefs" simultaneously possessed by a person – that is, of logically contradictory propositions simultaneously believed by a person. The following is one: *S* may believe "alcohol has all the toxicity of strychnine" while also believing "that a few drops of the spirits can have all the benefits of ambrosia" (translation: a little alcohol can make one feel good). In fact, the two propositions are not logically contradictory: that alcohol has the toxicity of strychnine is consistent with its being true that a little alcohol can make one feel good. Nor is there any logical contradiction in the propositions involved in similar examples of theirs: for instance, the combination of "*S* is not having an affair" with propositions constituting evidence (but not entailing) that *S* is having an affair.

Their "free love" example is unpersuasive for a different reason. **Brown & Kenrick** write, "we may be led to believe that 'free love' is a splendid idea while sexually aroused . . . and to believe precisely the opposite after viewing a film about AIDS." Now, surely, they do not want to claim that we never abandon any of our beliefs (there are many things I once believed that I no longer do). So why should we suppose that when the imagined people come to believe that free love is not a splendid thing, they still believe that it is a splendid thing? Furthermore, if they do simultaneously possess a relevant pair of beliefs about free love, one positive and the other negative, why should we maintain that the propositional contents are logically contradictory? Perhaps they believe that insofar as free love is pleasant, there is something to be said for it while also believing that since free love is very dangerous, there is much to be said against it. These two propositions are mutually consistent.

Kirsch contends that many hypnotized people acquire "contradictory beliefs."[1] He offers the following evidence. Many people "who displayed an apparent inability to bend an arm . . . indicated [1a] that they had tried to bend the arm and also [1b] that they could have bent their arm if they had really wanted to." Similarly, many "people who displayed suggested amnesia . . . claimed [2a] they wrote down every suggestion they could remember and [2b] that they could have remembered the suggestions if they really wanted to." Assuming that these people believed what they asserted, we again must ask whether the propositions believed are contradictory.

There is no logical contradiction in the conjunction of 1a and 1b. Consider an analogy. After losing a close tennis match, one might believe (3a) that one tried to win and (3b) that one would have (and hence could have) won if one had really wanted to win. There is no contradiction in this pair of propositions: indeed, one might *reasonably* believe that one would have won if one had tried considerably harder to win and that if one had been more strongly motivated ("really wanted") to win one would have tried a lot harder. Notice that 3a and 3b have the same form as 1a and 1b: since the former pair is not contradictory, neither is the latter pair. And although bending an arm normally is quite easy, the tennis analogy may not be far-fetched in the

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

present context. For these hypnotized individuals, arm-bending might have seemed to require a lot of effort, and more effort than they wanted to exert. The comparable data about amnesia can be handled along the same lines, although the quoted claims are less precise. (I take the subjects to mean that they wrote down every suggestion they consciously remembered, that they tried to remember, and that they could have remembered more suggestions if they had "really wanted" to.)

Dalgleish argues that "unexceptional cases of emotional self-deception . . . can involve holding two contradictory beliefs (p and $\sim p$) at the same time." He contends that "an individual can hold a propositional belief p while simultaneously having a higher-order emotional understanding of the situation consistent with $\sim p$." However, to claim that S has a higher-order understanding that is consistent with $\sim p$, or with S's believing that $\sim p$, is not yet to claim that S believes that $\sim p$. Presumably, many propositions *consistent* with – that is, not contradicted by – our emotional understandings of things are not believed by us. (That there is intelligent life on Mars does not contradict Al's emotional understanding of his mother's recent death, since that understanding has no bearing on Mars; but Al does not believe that there is intelligent life on Mars.) So Dalgleish must, and does, go further.

He contends that someone might believe that his brother is honest while also "having a sense that in fact he is deceitful." But does this "sense" amount to or encompass a *belief* that his brother is deceitful, or is it merely a suspicion that he is deceitful or a belief that there is evidence that he is deceitful (see sect. 4)? **Dalgleish** also claims that "everyday conversation" indicates that "paradoxical conflict" of the sort at issue is common (cf. **Dupuy, Losonsky,** and **Martin** on such conversation): people often say such things as "I know and believe that I'm a success at work because I only have to look at the evidence but deep down I still believe that I'm a failure." However, one must be careful in interpreting such assertions. **Barnden** plausibly takes claims of this kind to be metaphorical; and everyday conversation is influenced by everyday theories, many of which may be seriously misguided.

Sackeim & Gur, who have long advocated the dual belief model of self-deception, ask "How can information be experienced as threatening if there is no contradictory belief?" This question is answered in the target article (sect. 3). They contend that my criticism of their earlier work "can be interpreted as only admitting self-reports as evidence of beliefs" and, hence, as precluding "the study of evolutionary mechanisms for self-deception." However, my criticism is entirely consistent with the idea that actions and physiological tests can provide evidence of beliefs. It is Sackeim and Gur's tests and inferences that I called into question. Furthermore, in my own brief discussion of the study of evolutionary mechanisms for self-deception (sect. 4), I criticized a commitment to Sackeim and Gur's model of self-deception, not the enterprise itself. As **Heilman** explains, my position is in line with an evolutionary approach to investigating self-deception.

Evolution is an important consideration in **Lockard**'s commentary. Apparently, she takes me to have suggested that the question whether self-deception is an adaptive mechanism and related questions about the brain are somehow unimportant. However, the implications of my article for Lockard's preferred, important topics are not discouraging. If **Freeman** is right, my position on self-deception actually is supported by work on the limbic system.[2] If I am right, in studying self-deception Lockard probably is not studying a phenomenon that requires an intention to deceive oneself and satisfaction of the "dual belief" condition. That is bad news for her only if she is committed to the traditional conception of self-deception.

Like **Sackeim & Gur,** I believe that people "may be self-deceived about their loves, hates, wishes, and fears" (see Mele 1987a, pp. 157–58). Toward the end of their commentary, they discuss a situation of this kind. An individual – call him Zed – who is sexually attracted to X denies this but does "not simultaneously hold contradictory beliefs because the attraction to X is not believed." They observe – quite reasonably – that one should ask why "a mental content that is usually subject to awareness, such as a sexual attraction, in this case is not."[3] And without retracting the claim that Zed does not simultaneously hold contradictory beliefs, they suggest that "attribution of the strong form of self-deception may be warranted." This is puzzling, since "the strong form of self-deception," as they define it, requires the simultaneous presence of a belief that p and a belief that $\sim p$. Perhaps they are now thinking that what is crucial for strong self-deception is not "contradictory beliefs" but a certain kind of intentional agency. That issue is addressed in section R2.

Dupuy claims that my parting challenge "may imprison us in the sterile alternative: either the simultaneous presence of contradictory beliefs, or Mele's account." He does not explain why my account should be deemed sterile. Nor does he offer a third option. Dupuy writes: "One can . . . introduce a measure of individual self-deception along the lines of the interpersonal model. In one corner of his mind the man believes that F is CK, but in another corner he believes that F is not CK." But this directly appeals to one of the options – "contradictory beliefs" – in what he identifies as a sterile theoretical choice. (Although collective self-deception, which Dupuy mentions, is an interesting topic, it is beyond the scope of the target article and this response.)

It is noteworthy that none of the commentators has met my parting challenge. They have neither provided nor cited convincing evidence of the existence of instances of self-deception in which the self-deceiver simultaneously believes that p and believes that $\sim p$.

R2. Intentional self-deception? I argued that people may engage in intentional activities that contribute straightforwardly to their being self-deceived without having intentionally deceived themselves. Notice that (1) "S did A intentionally and S did B by doing A" does not entail (2) "S did B intentionally" (cf. sect. 5). Al intentionally flipped a light switch, and he thereby started an electrical fire; but Al did not intentionally start the fire. To take a case in which there is a strong statistical correlation between cause and effect, Bob intentionally punched Al's arm, thereby rupturing some capillaries there; but Bob did not intentionally rupture the capillaries. Setting aside some fine points, intentionally doing something X, as I and many others understand the notion, requires *aiming at X,* either as an end or as a means to (or constituent of) an end.[4] Al was not aiming at starting a fire, nor was Bob aiming at rupturing capillaries.

With this distinction in place, I can coherently claim, as I did (sects. 3, 5; cf. Mele 1987a, pp. 129–30), that people

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

may deceive themselves by, for example, intentionally ignoring unpleasant evidence or intentionally focusing on pleasant evidence without *intentionally deceiving* themselves. These people need not be aiming, consciously or unconsciously, at deceiving themselves, or at causing themselves to acquire or retain a certain belief. In the literature on the intentionality of self-deception, the question about the self-deceiver's *aim* is central. If there is a sense of "intentionally" – and elsewhere I have argued that there is no such legitimate sense (Mele & Sverdlik 1996) – in which Bob may be said to rupture capillaries in Al's arm intentionally, even though he was in no way (consciously or unconsciously) aiming at rupturing capillaries, perhaps much garden-variety self-deception is "intentional" in the same or a similar weak sense. But it is the stronger sense of "intentionally" that concerns me, the sense that entails "aiming at."

Martin's commentary seems to run these two senses of "intentionally" together (a legitimate and an illegitimate sense, if I am right). If his view is that "much self-deception" is intentional in the weak sense, it is consistent with mine. If his claim is that much self-deception is intentional in the stronger sense, he needs an argument for *that* claim. I granted that there may be unconscious intentions to deceive oneself, or to produce a certain belief in oneself (sect. 6), and I grant that there may be prereflective intentions to do these things. Martin has not shown that we have good reason to postulate such intentions in garden-variety instances of self-deception.

In **Perring**'s augmented version of Sam's case, does Sam intentionally deceive himself, or does he, without aiming at deceiving himself or at protecting his belief that Sally is faithful, intentionally avoid thinking about Sally's recent conduct and immerse himself in other activities, with the result that he retains his belief that Sally is faithful? This example is quite similar to my case of Beth, who, as "a consequence of [certain] intentional activities, . . . acquire[s] a false, unwarranted belief that her father cared more deeply for her than for anyone else" (sect. 5). I argued that what happens in a representative case of this kind is explicable independently of the supposition that the agents are aiming at deceiving themselves or at producing certain beliefs in themselves. Perring has not shown that cases of this kind are best interpreted as involving such aiming.

Bermudez argues that self-deception requires an intention to cause oneself to believe that *p* by "biasing [one's] cognitive processes," which intention is based on a desire to believe that *p* and a belief that "the best way to achieve this" is to bias one's cognitive processes in certain ways. He contends that this "intentionalist" view can answer two questions that my view cannot. (1) Why do some people who desire that *p* acquire a motivationally biased belief that *p* whereas other people, under very similar conditions, do not? (2) Why do some people who believe that *p* retain the belief owing to motivated biasing whereas others, under very similar conditions, revise the belief?

I address a version of the first question in the target article, identifying some relevant considerations, but offering no general answer and no account of conditions that are *causally sufficient* for the production of a motivationally biased belief (sect. 6 and n. 35). Given the similarity of the two questions, my remarks there apply to the latter question as well. On this complicated issue, **Bermudez** is much bolder than I. He claims that a desire to believe that *p* is present whenever people acquire a motivationally biased belief that *p* and absent otherwise, and that an intention to cause oneself to believe that *p* is present whenever a belief that *p* is retained owing to motivated biasing and absent whenever the belief is appropriately revised in the light of the evidence.[5] However, it is unlikely that everyone who desires to believe that *p* ends up believing that *p*, and Bermudez's claim about intention is subject to the objection raised in the target article (sect. 6) against a similar claim by Talbott. Bermudez's questions are important, but his answers are simplistic: not all desires and intentions are effective in producing their objects. Whether *S*'s desire that *p* will issue in (i.e., make a significant causal contribution to) *S*'s believing that *p* depends upon other facts about *S* and *S*'s circumstances. The same is true of an alleged desire to believe that *p* and an alleged intention of the kind Bermudez mentions.

Talbott contends that although "self-deception is not intentional *deception*," self-deceivers typically "try to bias their cognitive processes in favor of *p*." He observes that provided that "subjects have a choice between biasing their cognitive processes in favor of *p* or not biasing them," a Bayesian model would predict such trying. My position is that, typically, this is not a matter of choice and there is no need to suppose it is to explain the data. Return to Beth. Occasionally, she intentionally focuses her attention on certain pleasant memories, intentionally lingers over certain pleasant pictures, and intentionally turns her attention away from unpleasant memories of being left behind by her father. This behavior, which is utterly intelligible in light of hedonic considerations, makes a significant causal contribution to her acquiring a false, unwarranted belief that her father cared more deeply for her than for anyone else. Beth's intentional cognitive activities are explained, in part, by the attractiveness for her of the hypothesis that her father loved her most: if that hypothesis had been significantly less attractive, the hedonic difference between attention to memories and pictures that support it and attention to memories and pictures featuring her brothers in the spotlight of her father's affection would not have been so large. But there is no explanatory need to suppose that she was *trying* to bias her cognitive processes in favor of believing that hypothesis or that she *chose* to bias them in this direction. I have no objection, in principle, to unconscious trying. Indeed, I appeal to it in discussing Quattrone and Tversky's (1984) study. But where there is no explanatory need to postulate it and where other processes that are relatively well understood provide a straightforward explanation of the data, we do well to eschew appealing to such trying (cf. **Friedrich**).

I am happy to grant that people may intentionally deceive themselves, as section 6 of the target article makes plain. In Mele 1987a, drawing on Pascal, I describe a more realistic case in which an unhappy atheist, convinced that he would be much better off believing in God, consciously sets out to cause himself to believe that God exists by attending religious services, associating with religious people, and the like (pp. 133–34). Assume the following: he eventually succeeds; there is no God; and his evidence provides greater warrant for God's nonexistence than for God's existence. Then, I say, this agent deceives himself and is self-deceived. **Gibbins** discusses cases of a similar kind, but I fail to see how his suggestion that one can believe that *p* to a degree less than 0.5 helps resolve apparent problems

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

about such cases. Indeed, the suggestion itself is problematic. Why is allegedly believing that *p* to a degree of 0.1, say, to be counted as believing that *p*, as opposed to believing that $\sim p$ to a high degree? (Of course, we can believe that the probability of *p*'s being true is 0.1; but such a belief is not a belief that *p*. If it were, weather reports would frequently lead me to believe that it is going to rain in my town on a given day and simultaneously to believe that this is not going to happen.)

R3. Conceptually sufficient conditions for self-deception. Several commentators challenged my statement of conceptually sufficient conditions for entering self-deception in acquiring a belief (sect. 3). **Audi,** in an elegant commentary, offers an example that allegedly satisfies my conditions even though it is not a case of self-deception. My reply is that the causal process he describes does not satisfy the condition requiring that the causal connection between biasing and belief acquisition be nondeviant (condition 3). A deviant causal connection between an *X* and a *Y* is deviant relative to "normal" causal routes from *X*s to *Y*s. Here the relevant *X*s and *Y*s are, respectively, motivationally biased treatments (including gatherings) of data (at least seemingly) relevant to the truth value of *p* and acquisitions of beliefs that *p*. Audi has sketched an abnormal, hence deviant, route from an event of the first kind to an event of the second kind.

Audi and others argue that my conditions are insufficient for self-deception because they do not capture the "tension" that is *necessary* for self-deception. As Audi understands this tension, it "is ordinarily represented . . . by an avowal of *p* . . . *coexisting* with knowledge or at least true belief that not-*p*." In this way, Audi avoids what I dubbed the "static" puzzle (sect. 2): in his account, people who are self-deceived regarding *p* might avow *p*, but they do not actually believe that *p*. I have criticized this way of avoiding the puzzle elsewhere (Mele 1982; 1987b). In the target article, I argued that in garden-variety cases the person who is self-deceived regarding *p lacks* the true belief that $\sim p$. If that is right, the particular tension that Audi identifies is not part of garden-variety self-deception.

Losonsky identifies an alternative species of tension allegedly required for self-deception: self-deceivers have the unwarranted, false belief that *p* and lack the true belief that $\sim p$, but they possess evidence for $\sim p$ that is "active" in their "cognitive architecture," and this activity is manifested, for example, in recurrent or nagging doubts. He uses the claim that self-deception conceptually requires such conflict to support a distinction between self-deception and instances of "prejudice" or "bias" that satisfy the quartet of conditions I offered as conceptually sufficient for entering self-deception. **Martin** mentions a similar tension, "a cognitive conflict" such as "suspecting *p* and believing $\sim p$." And **Bach** contends that self-deception requires actively avoiding or suppressing certain thoughts, or ridding oneself of these thoughts when they occur.

Now, the set of conceptually sufficient conditions I offered certainly does not entail that there is no tension in self-deception. Nor did I claim that self-deception normally is tension-free. Satisfying my four conditions might *often* involve considerable psychic tension. The present question is whether any of the alleged kinds of tension is conceptually *necessary* for self-deception. And the answer is *no*. Even if Don, for example, is free of psychic conflict in the process of acquiring the belief that his article was unjustly rejected, he is self-deceived in acquiring that belief. The same is true of bigots who, without psychic conflict, satisfy my four conditions in acquiring a bigoted belief that *p*.

Losonsky suggests that I might be blind to the tension he describes because I have "a notion of a very stable and unified self" (cf. **Schmidt**). **Krebs et al.** make a similar suggestion about "contradictory beliefs." However, as I just observed, I have no quarrel with the idea that there is psychic tension: my claim is that such tension is not a conceptual requirement for self-deception. And my replies to various attempts to show that the "dual belief" condition sometimes is satisfied in self-deception do not rest on any particular conception of "the self."

My primary concern in the target article was the phenomenon of *entering* self-deception. Some commentators are more interested in the dynamics of *maintaining* self-deception, an important topic that I plan to address on another occasion. The processes **Bach** mentions, some of which were mentioned in the target article, are well suited to the maintenance of self-deception. However, Bach omits something crucial to self-deception that is present in my account. Some people who know that they have stomach problems or that their spouses are behaving strangely avoid thinking about these things. But this avoidance itself is not sufficient for self-deception; for they might not be deceived about any relevant proposition.

Gergen argues that the target article is not about self-deception. He claims that "common cultural sense," subjects in various experiments I reviewed, the experimenters themselves, and psychoanalysts are all in agreement about this; they all accept the traditional definition of self-deception that I rejected. Of course, I argued that acceptance of this definition rests on some mistakes, including, often, the mistaken theoretical idea that what happens in garden-variety self-deception can only, or best, be explained on the hypothesis that the "dual belief" and intention conditions are satisfied.[6] Although I take common cultural sense seriously when it intuitively identifies unanalyzed vignettes as cases of self-deception, I take it much less seriously when it offers a theory about the mechanisms at work in those cases. I doubt that common sense theories about complicated psychological or philosophical matters are likely to be much more successful than common sense theories in physics or chemistry. As for psychoanalysts, they offer evidence for their theories about what happens in these cases, and I and others can assess it. Furthermore, whether those who have investigated motivated bias would describe some instances of what they are investigating as self-deception will depend upon what they think self-deception is. That they think self-deception entails the "dual belief" and intention conditions, if indeed they do, certainly does not settle the matter. They may be, in **Bornstein**'s words, "clinging to outmoded (but familiar) ideas." (**Baumeister & Leith**'s commentary constitutes a nice counterexample to Gergen's claim that experimenters who study motivated bias see no overlap between that and self-deception.) And surely Quattrone and Tversky thought they were investigating self-deception.

Gergen asks whether there is "an empirically grounded difference between motivation and intention." Intentions are motivational attitudes, but they differ conceptually from other motivational attitudes – for example, mere desires (Mele 1992a). There certainly are studies con-

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

ducted under the rubric "intention" by psychologists who do not identify the concept of intention with the more general concept of motivation (see, e.g., Halisch & Kuhl 1987; Heckhausen 1991; cf. Taylor & Gollwitzer 1995). Whether those studies are actually about intention depends upon what intention is.

R4. Self-deception and interpersonal deception. I argued that whereas "stereotypical" interpersonal deception is intentional and involves there being a time at which the deceiver believes that $\sim p$ and the deceived believes that p, self-deception typically lacks these features. This obviously does not commit me to rejecting all alleged similarities between interpersonal deception and self-deception (see **Dupuy, Martin, Schmidt**).

Rachlin & Frankel argue that self-deception and interpersonal deception are "isomorphic" with one another because they have similar functions. My assertion that self-deception is *not* isomorphic with stereotypical interpersonal deception (sect. 7) is precisely the assertion reported in the first sentence of the present section. Many of Rachlin & Frankel's claims about functions are simply irrelevant to my assertion. Further, their conception of self-deception is flawed. They contend, for example, that we are self-deceived when "our beliefs are contrary to the present state of affairs." Surely, that is wrong. Just now a friend calling from Paris told me that it is raining there. I believe, accordingly, that it is raining there now. Suppose he was lying to me. Then my belief is contrary to the relevant present state of affairs: I am deceived about the weather in Paris, but I plainly am not *self*-deceived about it. Rachlin & Frankel also claim that "when successful in its natural function, self-deception becomes veridical perception." This implies that self-deception about past events is never successful in its natural function. Deceiving oneself into believing that one's spouse has not had an affair or that one's children have not used drugs cannot "become veridical perception," given that the past cannot be changed. Assuming that self-deception has a natural function, one wonders why it cannot be satisfied in a very common range of cases.

Since **Rachlin & Frankel** offer no argument that self-deception has *the* natural function they claim it has, the issue is difficult to assess. The same is true of their claim that "Most of the time, when we act against our immediate interests because we believe we will bring about a higher good that does not currently exist, we are, by Mele's criterion, self-deceived." People might act against immediate interests in sticking to a diet, declining another shot of bourbon, continuing to do homework, and so on, in the interest of producing a future, higher good. Is their claim that my position entails that these people are self-deceived (I don't see how), or that cases of this kind are in the minority in the relevant class of cases?

R5. Desire, belief, and explanation. In a nice bit of irony, **Lazar** criticizes my position for embracing a bit of belief-desire lore that **Foss** criticizes my position for violating. As Foss observes, I claim that a desire that p can play a causal role in the production of a belief that p by, for example, enhancing the vividness of evidence for p. But, he contends, desires have "explanatory force" only in connection with beliefs that identify (apparent) means to the desires' satisfaction; so desires cannot do what I say they can. What Foss has done is to over-generalize from a reasonable theory about how desires contribute to *intentional* conduct to an unreasonable theory about the causal roles of desire in general. Recognizing that mistake should open one's eyes to the empirical evidence I cited (e.g., Kunda 1987; 1990) that desire sometimes plays the roles I ascribed it.

As **Foss** notes, the causal connections between desire and belief emphasized in the target article are between a desire that p and a belief that p and include no intervening instrumental belief about the satisfaction of a desire (cf. **Bermudez**). Yet **Lazar** writes, "Mele suggests that, in many cases, it is the desire to believe that, together with some instrumental belief, accounts for the formation of the irrational belief." This assertion is mysterious. The key to the mystery, I suspect, is an unwarranted generalization from part of my discussion of Quattrone and Tversky's experiment.

If Quattrone and Tversky's subjects had not understood what a shift in tolerance in a certain direction was supposed to indicate, they would have had no motivation to try to alter their tolerance on the second trial. I argued that this understanding can play a role in their behavior without their believing that they are trying to alter their tolerance, and, hence, without their believing that they are trying to alter it in order to produce evidence of a healthy heart (sect. 4). **Lazar** insists that this is incomprehensible. She asks: "if agents are said to be shifting their tolerance *for the reason* of desiring to hold the belief, how can it be true, at the same time, that they never recognize their actions as instances of shifting tolerance?" The answer really is quite simple: the agent might believe that her pain reports are issued for the same reasons on this trial as on the earlier one and not recognize what is actually motivating those reports.

R6. Kinds of self-deception. A list of dimensions is discussed by **Bornstein,** that might prove useful in classifying various kinds of self-deception and **Krebs et al.** develop some related ideas. Here I will address some specific kinds of self-deception mentioned by commentators.

As I observed in the target article, we sometimes "deceive ourselves into believing that p is true even though we would like p to be false" (n. 6). Here one finds what might be dubbed "twisted" self-deception, as opposed to the "straight" variety in which what self-deceivers believe is something they want to be true. **Dalgleish** sketches an attractive explanation of twisted cases that, as he says, parallels and complements my own explanation of straight cases. He suggests that just as a desire that one's spouse not have an affair can prime cognitive biases the operation of which makes it easier to believe that one's spouse is not so engaged, so can jealousy by similar priming promote the false and unwarranted belief that one's spouse *is* having an affair (in the absence of any desire for the spouse's infidelity). This idea is well worth exploring.

Friedrich's position on hypothesis testing implies another interesting account of twisted self-deception. Whereas for many people, perhaps, it may be more important to avoid falsely believing that one's spouse is having an affair than to avoid falsely believing that one's spouse is not so engaged, the converse may well be true of some jealous people. Avoiding falsely believing that their spouses are faithful may be so important to certain jealous people that data suggestive of infidelity are particularly salient for them and contrary data quite pallid by comparison. This explanatory hypothesis is consistent with **Dalgleish**'s, but unlike

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

the latter it is suggested by a general model of pragmatic reasoning.

Friedrich presents his commentary as a cautiously friendly one. It is even friendlier than one might recognize. He challenges the distinction between "accuracy-driven and motivationally biased processing." My position on self-deception does not depend on there being any *actual* instances of accuracy-driven processing. Even if all actual hypothesis testing is motivated by an interest in minimizing costly mistakes, as Friedrich suggests, *norms* with respect to which bias is measured may be derived from an account of idealized, truth-seeking hypothesis testing. It may be that, relative to these norms, self-deceptive hypothesis testing is generally significantly more biased than non-self-deceptive hypothesis testing driven by an interest in avoiding costly mistakes. (Notice that bias comes in different degrees. A hypothesis tester's reasonably giving a loved one the benefit of the doubt, as we say, is distinguishable – at least in degree of bias – from more severe cases of bias discussed in the target article.)

Ainslie, like **Friedrich,** advocates a motivational account of hypothesis testing. I have examined his bold, motivational account of belief elsewhere (Mele 1993) and will pass over it here. I do wonder, however, how he would defend his claim that when "people 'deceive themselves,' they invariably seem to be discerning more occasion for good feeling (or less for bad) than a disinterested observer would" against the claim that there are cases of "twisted" self-deception.

Gorassini argues that there is a kind of self-deception that is "more intentional" than the kind to which I devote most of the target article. He appeals to the phenomenon of acquiring a belief that one has a certain desired property as a consequence of *acting as if* one has that property and to data about game-playing and hypnosis. I addressed the first phenomenon in Mele 1987a, arguing that representative self-deceptive instances of it are plausibly explained on my own model of self-deception (pp. 151, 157–58). Given constraints on space, I forego further discussion of it here. Regarding game-playing, a distinction between deceiving oneself and *pretending* is in order; I am not convinced that there is self-deception in these cases. Gorassini's discussion of hypnosis, like **Kirsch**'s, is instructive. But nothing in the target article commits me to denying that intentionally playing along with one's hypnotist in order to experience hypnosis can contribute to one's acquiring such false beliefs as that one is not voluntarily moving one's arm (see sect. R2).

Bornstein, Dupuy, and **Schmidt** observe that the etiology of at least some instances of self-deception has, or might have, an important social dimension. I agree. For example, as I noted elsewhere (Mele 1987a, p. 150; cf. p. 158), a desire that *p* "may lead via social routes to nonintentional selective exposure to data" supportive of *p*. An older boy who wants it to be true that he is a natural leader but lacks the respect of his peers may find the company of younger teenagers more congenial, and his hedonically motivated "choice of 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 entering self-deception in acquiring certain beliefs about his leadership abilities. A careful examination of social routes to personal self-deception promises to prove fruitful. Reasonable constraints on space placed the issue beyond the scope of the target article, and the same is true of this response.

R7. Additional conceptual issues. The idea that self-deception is a "historical concept" is rejected by **Audi**. He writes: "If I am self-deceived, so is my perfect replica at the very moment of his creation." I disagree. Some concepts, including some psychological ones, are historical in Audi's sense, and I take self-deception to be among them. Consider the concept "remembering" (as opposed to "seeming to remember"). My perfect replica at the moment of his creation does not remember gaining employment at my college; one cannot (actually) remember something that has never happened. Similarly, as I understand self-deception, beings who have not deceived themselves are not self-deceived, no matter what else is true of them.

Johnson, is his thoughtful commentary, argues that "definitive ascriptions of self-deception in everyday life are simply not to be had" and that this "may well lie in the phenomenon" itself rather than in my proposed set of conceptually sufficient conditions for self-deception. As he observes, I did not claim that we can be certain who is self-deceived and who is not. But it is worth pointing out that we can have *significant evidence* that my conditions are satisfied in particular cases. We can have good grounds for attributing to *S* a belief that we know to be false. By carefully studying *S*, we can learn a lot about what relevant data *S* possesses. And we can construct tests to provide evidence of motivated bias. For example, we can give impartial subjects what we are fairly confident are *S*'s data and ask them what conclusion is most strongly supported by these data, and we can test *S*'s ability to make inferences about other matters in which *S* has little or no motivational stake. Of course, Johnson need not disagree with any of this, given his concern with *definitive* ascriptions. And I agree with him that "a full understanding of real self-deception" will include an understanding of the practice of ascribing self-deception.

Johnson does not present his reasonable worries about the possibility of definitive attributions of self-deception to individuals as grounds for denying that self-deception occurs. And rightly so. As **Baumeister & Leith** observe, our best evidence for the existence of self-deception, as I (and they and Johnson) conceive the phenomenon, is obtained by aggregation, and I warmly welcome what they describe as an extension of my analysis.

NOTES

1. He also suggests that they may be "unaware of . . . the discrepancy between the two beliefs" (cf. **Bermudez**).

2. I should add that **Freeman** and I apparently differ somewhat, both in how we conceive of intentional action and in our interpretation of some of Libet's data. For all Libet has shown, the mental item that appears on the unconscious scene about 550 msec before muscular motion begins in the scenarios he studies might be an urge or desire that antedates an actual decision or intention. But that is another story.

3. See my discussion of Quattrone and Tversky's study in the target article (sect. 4) and in section R5. Notice that what one is trying to do "is usually subject to awareness."

4. For a detailed conceptual analysis of intentional action, see Mele & Moser 1994.

5. Presumably, **Bermudez** means that these agents intend to cause themselves to *continue* believing that *p*.

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

6. Some theorists accept the definition primarily on linguistic grounds. These grounds were challenged in the target article (sect. 2).

References

Letters "a" and "r" appearing before authors' initials refer to target article and response respectively.

Ainslie, G. (1992) *Picoeconomics: The strategic interaction of successive motivational states within the person.* Cambridge University Press. [aARM, GA]

(1993) A picoeconomic rationale for social constructionism. *Behavior and Philosophy* 21:63–75. [GA]

(1994) Is rationality just a bookkeeping system? Address to the American Philosophical Association, April 2, 1994, Los Angeles, CA. [GA]

Aristotle (1991) *The art of rhetoric.* Penguin. [TD]

Audi, R. (1985) Self-deception and rationality. In: *Self-deception and self-understanding,* ed. M. Martin. University Press of Kansas. [aARM, MWM]

(1989) Self-deception and practical reasoning. *Canadian Journal of Philosophy* 19:247–66. [aARM]

(in press) Self-deception, rationalization and the ethics of belief: An essay in moral psychology. In: *Moral knowledge and ethical character,* ed. R. Audi. Oxford University Press. [RA]

Aumann, B. (1976) Agreeing to disagree. *Annals of Statistics* 4:1236–39. [J-PD]

Bach, K. (1981) An analysis of self-deception. *Philosophy and Phenomenological Research* 41:351–70. [aARM, KB]

(1984) Default reasoning. *Pacific Philosophical Quarterly* 65:37–58. [KB]

(1992) Review of *Perspectives on self-deception,* ed. B. McLaughlin & A. O. Rorty. *Noûs* 26:495–504. [KB]

(1994) Emotional disorder and attention. In: *Philosophical psychopathology,* ed. G. Graham & L. Stephens. MIT Press. [KB]

Barlow, J. (1989) *Darwin, sex and status.* University of Toronto Press. [JSL]

Barnard, P. (1985) Interacting cognitive subsystems: A psycholinguistic approach to short-term memory. In: *Progress in the psychology of language,* vol. 2., ed. A. Ellis. Erlbaum. [TD]

Barnden, J. A. (1996) Consciousness and common-sense metaphors of mind. In: *Reaching for mind: Foundations of cognitive science,* ed. S. O'Nuallain, P. McKevitt & E. MacAogain. John Benjamin. [JAB]

Barnden, J. A., Helmreich, S., Iverson, E. & Stein, G. C. (1996) Artificial intelligence and metaphors of mind: Within-vehicle reasoning and its benefits. *Metaphor and Symbolic Activity* 11(2):101–23. [JAB]

Baron, J. (1988) *Thinking and deciding.* Cambridge University Press. [aARM]

Barrie, J. M., Freeman, W. J. & Lenhart, M. (1996) Modulation by discriminative training of spatial patterns of gamma EEG amplitude and phase in neocortex of rabbits. *Journal of neurophysiology*: July, 76:520–539. [WJF]

Baumann, A. O., Deber, R. B. & Thompson, G. G. (1991) Overconfidence among physicians and nurses: The micro-certainty, macro-uncertainty phenomenon. *Social Science and Medicine* 32:167–74. [MH]

Baumeister, R. F. (in press) The self. In: *Handbook of social psychology,* 4th ed., ed. D. T. Gilbert, S. T. Fiske & G. Lindzey. McGraw-Hill. [RFB]

Baumeister, R. & Cairns, K. (1992) Repression and self-presentation: When audiences interfere with self-deceptive strategies. *Journal of Personality and Social Psychology* 62:851–62. [aARM]

Baumeister, R. F. & Newman, L. S. (1994) Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin* 20:3–19. [RFB]

Bittner, R. (1988) Understanding a self-deceiver. In: *Perspectives on self-deception,* ed. B. P. McLaughlin & A. O. Rorty. University of California Press. [JAB]

Bornstein, R. F. (1991) Manuscript review in psychology: Psychometrics, demand characteristics, and an alternative model. *Journal of Mind and Behavior* 12:429–68. [RFB]

Bourdieu, P. (1977) *Outline of a theory of practice,* trans. Richard Nice. Cambridge University Press. [J-PD]

Brown, J. D. & Dutton, K. A. (1995) Truth and consequences: The costs and benefits of accurate self-knowledge. *Personality and Social Psychology Bulletin* 21:1288–96. [JF]

Burley, N. (1979) The evolution of concealed ovulation. *American Naturalist* 114:835–58. [MH]

Butler, G. & Mathews, A. (1983) Cognitive processes in anxiety. *Advances in Behavioral Research and Therapy* 5:51–62. [TD]

Butler, G. & Mathews, A. (1987) Anticipatory anxiety and risk perception. *Cognitive Therapy and Research* 91:551–65. [TD]

Butler, J. (1896) Upon self-deceit. In: *The works of Joseph Butler,* vol. 2, ed. W. Gladstone. Clarendon Press. [MWM]

Campion, J., Latto, R. & Smith, Y. (1983) Is blindsight an effect of scattered light, spared cortex, and near-threshold vision? *Behavioral and Brain Sciences* 6:423–86. [aARM]

Cialdini, R. B., Kallgren, C. A. & Reno, R. R. (1991) A focus theory of normative conduct. *Advances in Experimental Psychology* 24:201–34. [SLB]

Clancey, W. J. (1993) Situated action: A neuropsychological interpretation response to Vera and Simon. *Cognitive Science* 17:87–116. [WJF]

Comey, G. & Kirsch, I. (1995) Intentional and spontaneous imagery in hypnosis. Paper to the Society for Clinical and Experimental Hypnosis, San Antonio, TX (November 1995). [IK]

Davidson, D. (1982) Paradoxes of irrationality. In: *Philosophical essays on Freud,* ed. R. Wollheim & J. Hopkins. Cambridge University Press. [aARM]

(1985) Deception and division. In: *Actions and events,* ed. E. LePore & B. McLaughlin. Blackwell. [aARM, JAB, ML]

(1986) Deception and division. In: *The multiple self,* ed. J. Elster. Cambridge University Press. [AL]

Dennett, D. (1987) *The intentional stance.* MIT Press. [CTS]

(1991) *Consciousness explained.* Little, Brown & Co. [ML]

Dewey, J. (1914) Psychological doctrine in philosophical teaching. *Journal of Philosophy* 11:505–12. [WJF]

Dollard, J. & Miller, N. (1950) *Personality and psychotherapy: An analysis in terms of learning, thinking, and culture.* McGraw-Hill. [SLB]

Douglas, W. & Gibbins, K. (1983) Inadequacy of voice recognition as a demonstration of self-deception. *Journal of Personality and Social Psychology* 44:589–92. [aARM]

Dupuy, J.-P. (1995) Not to know what one knows: Some paradoxes of self-deception. *Diogenes* 169:53–68. [J-PD]

Ekman, P. (1985) *Telling lies.* Norton. [JSL]

Engel, P. (1992) *Etats d'esprit: Questions de philosophie de l'esprit.* Alinéa. [CTS]

Festinger, L. (1957) *A theory of cognitive dissonance.* Stanford University Press. [aARM]

(1964) *Conflict, decision, and dissonance.* Stanford University Press. [aARM]

Festinger, L., Riecken, H. W. & Schachter, S. (1964) *When prophecy fails.* Harper & Row. [KG]

Fingarette, H. (1969) *Self-deception.* Humanities Press. [aARM, JSL, MWM, HAS]

Fleming, J. H. & Darley, J. M. (1991) Mixed messages: The multiple audience problem in strategic communication. *Social Cognition* 9:25–46. [RFB]

Foss, J. (1976) A rule of minimal rationality: The logical link between beliefs and values. *Inquiry* 19:341–53. [JEF]

(1980) Rethinking self-deception. *American Philosophical Quarterly* 17:237–43. [JEF]

Freeman, W. J. (1975) *Mass action in the nervous system.* Academic Press. [WJF]

(1992) Tutorial in neurobiology. *International Journal of Bifurcation and Chaos* 2:451–82. [WJF]

(1995) *Societies of brains. A study in the neuroscience of love and hate.* Erlbaum. [WJF]

Freud, S. (1915/1956) Instincts and their vicissitudes. In: *The standard edition of the complete psychological works of Sigmund Freud,* vol. 14, ed. J. Strachey & A. Freud. Hogarth. [GA]

Frey, D. (1986) Recent research on selective exposure to information. In: *Advances in experimental social psychology,* vol. 19, ed. L. Berkowitz. Academic Press. [aARM]

Friedrich, J. (1993) Primary error detection and minimization (PEDMIN) strategies in social cognition: A reinterpretation of confirmation bias phenomena. *Psychological Review* 100:298–319. [JF]

Gardiner, P. (1969–70) Error, faith and self-deception. *Aristotelian Society Proceedings* 70:221–43. [MWM]

Gardner, S. (1993) *Irrationality and the philosophy of psychoanalysis.* Cambridge University Press. [AL]

Gergen, K. (1985) The ethnopsychology of self-deception. In: *Self-deception and self-understanding,* ed. M. Martin. University Press of Kansas. [aARM, KJG]

(1988) If persons are texts. In: *Hermeneutics and psychological theory,* ed. S. B. Messer, L. A. Sass & R. L. Woolfolk. Rutgers University Press. [KJG]

(1994) *Realities and relationships.* Harvard University Press. [KJG]

Gibbins, K. & Douglas, W. (1985) Voice recognition and self-deception: A reply to Sackeim and Gur. *Journal of Personality and Social Psychology* 48:1369–72. [aARM]

Gibson, J. J. (1979) *The ecological approach to visual perception.* Houghton Mifflin. [WJF]

Gilovich, T. (1991) *How we know what isn't so.* Macmillan. [aARM, GA, KG]

Goffman, E. (1959) *The presentation of self in everyday life.* Doubleday. [MH]

Gorassini, D. R. (in press a) Hypnotic responsiveness: A cognitive-behavioral

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

analysis of self-deception. In: *Clinical hypnosis and self-regulation therapy: A cognitive-behavioral perspective,* ed. I. Kirsch, A. Capafons, S. Amigó & E. Cardeña. American Psychological Association Books. [DRG, IK]
(in press b) Strategy selection and hypnotic performance. *Contemporary Hypnosis.* [DRG]
Gorassini, D. R. & Spanos, N. P. (1986) A cognitive-social skills approach to the successful modification of hypnotic susceptibility. *Journal of Personality and Social Psychology* 50:1004–12. [DRG]
Green, L., Fry, A. & Myerson, J. (1994) Discounting of delayed rewards: A life-span comparison. *Psychonomic Science* 5:33–36. [GA]
Greenwald, A. (1988) Self-knowledge and self-deception. In *Self-deception: An adaptive mechanism?* ed. J. Lockard & D. Paulhus. Prentice-Hall. [aARM]
Greenwald, A. G. & Banaji, M. R. (1995) Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review* 102:4–27. [RFB]
Greenwald, A. S. (1988) Self-knowledge and self-deception. In: *Self-deception: An adaptive mechanism?* ed. J. S. Lockard & D. L. Paulhus. Prentice Hall. [HAS]
Griffen, D. W., Dunning, D., Ross, L. (1990) The role of construal processes in overconfident predictions about the self and others. *Journal of Personality and Social Psychology* 59:1128–39. [MH]
Gur, R. & Sackeim, H. (1979) Self-deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology* 37:147–69. [aARM, TD, KG, MH, ML, HAD]
Haight, M. (1980) *A study of self-deception.* Harvester Press. [aARM]
Halisch, F. & Kuhl, J. (1987) *Motivation, intention, and volition.* Springer-Verlag. [rARM]
Heckhausen, H. (1991) *Motivation and action.* Springer-Verlag. [rARM]
Higgins, R., Snyder, C. & Berglas, S. (1990) *Self-handicapping: The paradox that isn't.* Plenum Press. [aARM]
Hilgard, E. R. (1965) *Hypnotic susceptibility.* Harcourt, Brace & World. [IK]
(1986) *Divided consciousness: Multiple controls in human thought and action,* expanded ed. Wiley. [IK]
Hirt, E. R., Deppe, R. K. & Gordon, L. J. (1991) Self-reported versus behavioral self-handicapping: Empirical evidence for a theoretical distinction. *Journal of Personality and Social Psychology* 61:981–91. [JF]
Jacoby, L. L., Toth, J. P., Lindsay, S. D. & Debner, J. A. (1992) Lectures for a layperson: Methods for revealing unconscious processes. In: *Perception without awareness: Cognitive, clinical and social perspectives,* ed. R. F. Bornstein & T. S. Pittman. Guilford Press. [RFB]
Jacques, F. (1979) *Dialogiques, Recherches logiques sur le dialogue.* PUF. [CTS]
(1985) *L'Espace logique de l'interlocution.* PUF. [CTS]
(1990) De *On denoting* de B. Russell à *On referring* de P. F. Strawson, l'avenir d'un paradigme. In: *Hermès VII: Bertrand Russell, de la logique à la politique.* Editions du CNRS. [CTS]
Johnson, E. A. (1996) Children's understanding of epistemic conduct in self-deception and other false belief stories. Submitted manuscript. [EAJ]
Johnson, M. K. & Multhaup, K. S. (1992) Emotion and MEM. In: *The handbook of emotion and memory: Research and theory,* ed. S.-A. Christianson. Erlbaum. [TD]
Johnston, M. (1988) Self-deception and the nature of mind. In: *Perspectives on self-deception,* ed. B. McLaughlin & A. O. Rorty. University of California Press. [aARM, JAB, KB]
Jones, S. C. (1973) Self and interpersonal evaluations: Esteem theories versus consistency theories. *Psychological Bulletin* 79:185–99. [KG]
Kimura, D. (1987) Are men's and women's brains really different? *Canadian Journal of Psychology* 28:133–47. [JSL]
King-Farlow, J. (1963) Self-deceivers and Sartrian seducers. *Analysis* 23:131–36. [JAB]
Kipp, D. (1980) On self-deception. *Philosophical Quarterly* 30:305–17. [aARM]
Kirby, K. N. & Herrnstein, R. J. (1995) Preference reversals due to myopic discounting of delayed reward. *Psychological Science* 6:83–89. [GA]
Kirsch, I. (1985) Response expectancy as a determinant of experience and behavior. *American Psychologist* 40:1189–1202. [IK]
(in press) Specifying nonspecifics: Psychological mechanisms of placebo effects. In: *Just a placebo? An interdisciplinary exploration,* ed. A. Harrington. Harvard University Press. [IK]
Kirsch, I. & Lynn, S. J. (1995) The altered state of hypnosis: Changes in the theoretical landscape. *American Psychologist* 50:846–58. [IK]
Kunda, Z. (1987) Motivated inference: Self-serving generation and evaluation of causal theories. *Journal of Personality and Social Psychology* 53:636–47. [aARM, GA]
(1990) The case for motivated reasoning. *Psychological Bulletin* 108:480–98. [arARM, GA, RFB, JF]
Lazar, A. (forthcoming a) Deceiving oneself or self-deceived? On the formation of beliefs "Under the Influence." [AL]
(forthcoming b) Division and deception: On Davidson's theory of self-deception. In: *Theories on self-deception,* ed. J. P. Dupuy. CSLI Press. [AL]
Lewicka, M. (1989) Toward a pragmatic perspective on cognition: Does evaluative meaning influence rationality of lay inferences? *Polish Psychological Bulletin* 20:267–85. [JF]
(1992) Pragmatic reasoning schemata with differing affective value of a consequent of logical implication. *Polish Psychological Bulletin* 23:237–52. [JF]
Lewinsohn, P. M., Mischel, W., Chaplin, W. & Barton, R. (1980) Social competence and depression: The role of illusory self-perceptions. *Journal of Abnormal Psychology* 89:203–12. [GA]
Lewis, B. L. (1996) Self-deception: A postmodern reflection. *Journal of Theoretical and Philosophical Psychology* 16:49–65. [KJG]
Lewis, D. K. (1969) *Convention: A philosophical study.* Harvard University Press. [J-PD]
Libet, B. (1994) *Neurophysiology of consciousness: Selected papers and new essays.* Birkhauser. [WJF]
Lockard, J. (1978) Commentary: Speculations on the adaptive significance of cognition and consciousness. *Behavioral and Brain Sciences* 4:583–84. [JSL]
(1980) Speculations on the adaptive significance of self-deception. In: *The evolution of human social behavior,* ed. J. Lockard. Elsevier. [JSL]
(1988) Origins of self-deception: Is lying to oneself uniquely human? In: *Self-deception: An adaptive mechanism?* ed. J. Lockard & D. Paulhus. Prentice-Hall. [JSL]
Lockard, J. & Mateer, C. (1988) Neural bases of self-deception. In: *Self-deception: An adaptive mechanism?* ed. J. Lockard & D. Paulhus. Prentice-Hall. [JSL]
Lockard, J. & Paulhus, D. (1988) *Self-deception: An adaptive mechanism?* Prentice-Hall. [aARM, JSL]
Lyons, W. (1980) *Emotion.* Cambridge University Press. [TD]
Martin, M. (1985) *Self-deception and self-understanding.* University Press of Kansas. [aARM]
Martin, M. (1986) *Self-deception and morality.* University Press of Kansas. [MWM]
Martindale, C. (1980) Subselves. In: *Review of personality and social psychology,* ed. L. Wheeler. Sage. [SLB]
(1991) *Cognitive psychology: A neural-network approach.* Brooks/Cole. [SLB]
Mateer, C., Poler, S. & Ojemann, G. (1982) Sexual variation in cortical localization of naming as determined by stimulation mapping. *Behavioral and Brain Sciences* 5:310–11. [JSL]
McConkey, K. M. (1991) The construction and resolution of experience and behavior in hypnosis. In: *Theories of hypnosis: Current models and perspectives,* ed. S. J. Lynn & J. W. Rhue. Guilford Press. [IK]
McGlone, J. (1980) Sex differences in human brain asymmetry: A critical survey. *Behavioral and Brain Sciences* 5:215–64. [JSL]
McLaughlin, B. (1988) Exploring the possibility of self-deception in belief. In: *Perspectives on self-deception,* ed. B. McLaughlin & A. O. Rorty. University of California Press. [aARM, KB]
Mele, A. (1982) Self-deception, action and will: Comments. *Erkenntnis* 18:159–64. [rARM]
(1983) Self-deception. *Philosophical Quarterly* 33:365–77. [aARM]
(1987a) *Irrationality.* Oxford University Press. [arARM]
(1987b) Recent work on self-deception. *American Philosophical Quarterly* 24:1–17. [arARM, JAB]
(1992a) *Springs of action.* Oxford University Press. [arARM, JLB]
(1992b) Recent work on intentional action. *American Philosophical Quarterly* 29:199–217. [aARM]
(1993) Motivated belief. *Behavior and Philosophy* 21:19–27. [rARM]
(1995) *Autonomous agents: From self-control to autonomy.* Oxford University Press. [aARM]
Mele, A. & Moser, P. (1994) Intentional action. *Noûs* 28:39–68. [arARM]
Mele, A. & Sverdlik, S. (1996) Intention, intentional action, and moral responsibility. *Philosophical Studies* 82:265–87. [rARM]
Merleau-Ponty, M. (1945/1963) *Phenomenology of perception,* translated by C. Smith. Humanities Press. [WJF]
Miller, E. M. (1996) Concealed ovulation as a strategy for increasing per capita paternal investment. *Mankind Quarterly* 37:297–333. [MH]
Mills, M. E. (1996) Evolutionary epochs in neural evolution: A model and some implications for human psychological functioning. Paper presented at a meeting of the Human Behavior and Evolution Society, Chicago, IL (June 1996). [SLB]
Minsky, M. (1985) *The society of mind.* Simon & Schuster. [ML]
Mitchell, R. W. & Thompson, N. S. (1986) *Deception, perspectives on human and nonhuman deceit.* State University of New York Press. [MH]
Nisbett, R. & Ross, L. (1980) *Human inference: Strategies and shortcomings of social judgment.* Prentice-Hall. [aARM]

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press

Oatley, K. & Johnson-Laird, P. N. (1987) Towards a cognitive theory of emotions. *Cognition and Emotion* 1:29–50. [TD]
Ojemann, G. (1979) A review of the neurological basis of human cognition, with special emphasis on language. *Allied Health and Behavioral Sciences* 1:341–84. [JSL]
Orne, M. T. (1959) The nature of hypnosis: Artifact and essence. *Journal of Abnormal Psychology* 58:277–99. [IK]
Paulhus, D. (1986) Self-deception and impression management in test responses. In: *Personality assessment via questionnaires,* ed. A. Angleitner & J. S. Wiggins. Springer. [MH]
Pears, D. (1984) *Motivated irrationality*. Oxford University Press. [aARM, AL]
(1991) Self-deceptive belief-formation. *Synthese* 89:393–405. [aARM]
Peele, S. (1989) *Diseasing of America: Addiction treatment out of control.* Lexington Books. [aARM]
Perris, C. (1988) The foundations of cognitive psychotherapy and its standing in relation to other psychotherapies. In: *Cognitive therapy: Theory and practice,* ed. C. Perris, I. M. Blackburn & H. Perris. Springer-Verlag. [GA]
Piaget, J. (1974/1980) *Adaptation and intelligence.* Hermann. [WJF]
Plato (1953) *Cratylus.* In: *The dialogues of Plato,* trans. B. Jowett. Clarendon Press. [aARM]
Popper, K. R. & Eccles, J. C. (1977) *The self and its brain.* Springer. [WJF]
Power, M. J. & Dalgleish, T. (1997) *Cognition and motion: From order to disorder.* Erlbaum. [TD]
Putnam, H. (1981) Brains in a vat. In: *Reason, truth and history,* ed. H. Putnam. Cambridge University Press. [CTS]
Quattrone, G. & Tversky, A. (1984) Causal versus diagnostic contingencies: On self-deception and on the voter's illusion. *Journal of Personality and Social Psychology* 46:237–48. [arARM, GA, HR]
Rorty, A. O. (1980) Self-deception, akrasia, and irrationality. *Social Science Information* 19: 905–22. [aARM]
(1988) The deceptive self: Liars, layers, and lairs. In: *Perspectives on self-deception,* ed. B. P. McLaughlin & A. O. Rorty. University of California Press. [JAB, AL]
Sackeim, H. A. (1983) Self-deception, self-esteem, and depression: The adaptive value of lying to oneself. In: *Empirical studies of psychoanalytic theory,* vol. 1, ed. J. Masling. The Analytic Press. [HAS]
(1988) Self-deception: A synthesis. In: Self-deception: *An adaptive mechanism?* ed. J. Lockard & D. Paulhus. Prentice-Hall. [aARM]
Sackeim, H. A. & Gur, R. (1978) Self-deception, self-confrontation, and consciousness. In: *Consciousness and self-regulation,* vol. 2, ed. G. Schwartz & D. Shapiro. Plenum Press. [aARM, TD, IK, MK, HAS]
(1979) Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology* 47:213–15. [JSL]
(1985) Voice recognition and the ontological status of self-deception. *Journal of Personality and Social Psychology* 48:1365–68. [aARM, TD, ML]
Sarbin, T. R. (1989) The constructions and reconstruction of hypnosis. In: *Hypnosis: The cognitive-behavioral perspective,* ed. N. P. Spanos & J. F. Chaves. Prometheus Books. [IK]
Sartre, J.-P. (1966) *Being and nothingness,* trans. H. E. Barnes. Simon & Schuster/Pocket Books. [J-PD, MWM]
Schmidt, C. (in press) The systemics of dialogism: On the prevalence of the self in HCI design. *Journal of the American Society for Information Science* [special issue on the "Human-Computer Interface"]. Wiley. [CTS]
Schmitt, F. F. (1988) Epistemic dimensions of self-deception. In: *Perspectives on self-deception,* ed. B. P. McLaughlin & A. O. Rorty. University of California Press. [EAJ]
Schrauger, S. (1975) Responses to evaluation as a function of initial self-perceptions. *Psychological Bulletin* 82:581–96. [KG]
Silver, M., Sabini, J. & Miceli, M. (1989) On knowing self-deception. *Journal for the Theory of Social Behaviour* 19:213–27. [aARM]
Sorensen, R. (1985) Self-deception and scattered events. *Mind* 94:64–69. [aARM]
Spanos, N. P. (1986) Hypnotic behavior: A social psychological interpretation of amnesia, analgesia, and "trance logic." *Behavioral and Brain Sciences* 9:449–52. [DRG, IK]
Spanos, N. P. & Gorassini, D. R. (1984) Structure of hypnotic test suggestions and attributions of responding involuntarily. *Journal of Personality and Social Psychology* 46:688–96. [DRG]
Sperry, R. W. (1982) Some effects of disconnecting the cerebral hemispheres (Nobel Lecture) *Science* 217:1223–26. [WJF]
Springer, S. & Deutsch, G. (1985) Left brain, right brain. W. H. Freeman. [JSL]
Swann, W. B., Jr. (1987) Identity negotiation: Where two roads meet. *Journal of Personality and Social Psychology* 53:1038–51. [DRG]
Szabados, B. (1985) The self, its passions, and self-deception. In: *Self-deception and self-understanding,* ed. M. Martin. University Press of Kansas. [aARM]
Talbott, W. J. (1995) Intentional self-deception in a single, coherent self. *Philosophy and Phenomenological Research* 55:27–74. [aARM, AL]
Taylor, S. E. (1989) *Positive illusions: Creative self-deception and the healthy mind.* Basic Books. [aARM, DRG, MH]
Taylor, S. E. & Brown, J. D. (1988) Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin* 10:193–210. [MH]
Taylor, S. E. & Fiske, S. (1975) Point of view and perceptions of causality. *Journal of Personality and Social Psychology* 32:439–45. [aARM]
(1978) Salience, attention and attribution: Top of the head phenomena. In: *Advances in experimental social psychology,* vol. 11, ed. L. Berkowitz. Academic Press. [aARM]
Taylor, S. E. & Gollwitzer, P. (1995) Effects of mindset on positive illusions. *Journal of Personality and Social Psychology* 69:213–26. [rARM]
Taylor, S. E. & Thompson, S. (1982) Stalking the elusive "vividness" effect. *Psychological Review* 89:155–81. [aARM]
Teasdale, J. & Barnard, P. (1993) *Affect, cognition and change.* Erlbaum. [TD]
Tooby, J. & Cosmides, L. (1992) The psychological foundations of culture. In: *The adapted mind: Evolutionary psychology and the generation of culture,* ed. J. H. Barkow, L. Cosmides & J. Tooby. Oxford University Press. [SLB]
Trivers, R. (1985) *Social evolution.* Benjamin/Cummings. [aARM, JSL]
Tversky, A. & Kahnemann, D. (1973) Availability: A heuristic for judging frequency and probability. *Cognitive Psychology* 5:207–32. [aARM]
Weiskrantz, L. (1986) *Blindsight: A case study and implications.* Oxford University Press.
Whisner, W. (1993) Self-deception and other-person deception: A new conceptualization of one central type of self-deception. *Philosophia* 22:223–40. [AL]
Williams, J. M. G., Watts, F., MacLeod, C. & Mathews, A. (1988) *Cognitive psychology and emotional disorders.* Wiley. [GA]
Wyer, R. S., Jr. & Frey, D. (1983) The effects of feedback about self and others on the recall and judgements of feedback-relevant information. *Journal of Experimental Social Psychology* 19:540–59. [JF]
Zeigler, H. (1964) Displacement activity and motivation theory: A case study in the history of ethology. *Psychological Bulletin* 61:362–76. [JSL]

https://doi.org/10.1017/S0140525X97440037 Published online by Cambridge University Press