# The Spandrels of Self-Deception: Prospects for a Biological Theory of a Mental Phenomenon

**D. S. Neil Van Leeuwen**

> Self-deception is so undeniable a fact of human life that if anyone tried to deny its existence, the proper response would be to accuse this person of it.
>
> —Allen Wood[1]

## 1. Introduction to the Problems

Even prior to a characterization of self-deception, it's easy to see that it's widespread. This is puzzling, because self-deception seems paradoxical. Here are three apparent problems. First, self-deception seems to involve a conceptual contradiction; in order to *deceive*, one must believe the contrary of the deception one is perpetrating, but if one believes the contrary, it seems impossible for *that very self* to believe the deception.[2] Second, a view of the mind has become widely accepted that makes rationality constitutive and exhaustive of the mental; this is the interpretive view of the mental put forth by Donald Davidson and followers.[3] According to this view, the idea of attributing irrational beliefs to agents makes little sense, for we cannot make sense of what someone's beliefs *are* unless they are rational. But self-deception is highly irrational.[4] Third, a brief survey of other cognitive mechanisms (such as vision) reveals that most are well suited for giving us reliable information about ourselves and our environment. Self-deception, however, undermines knowledge of ourselves and the world. If having good information brings fitness benefits to organisms, how is it that self-deception was not weeded out by natural selection? Otherwise put, why does self-deception exist? (For convenience, I shall refer to these as puzzles 1, 2, and 3.)

Puzzles such as these make self-deception an intriguing explanatory target. Philosophical literature on self-deception abounds. Much of the philosophical literature,

however, focuses on characterizing the phenomenon in a way that avoids the paradoxes while still making sense of the term "self-deception." One approach, favored by Davidson (1982, 1985, 1998) and Pears (1984), is to draw divisions between parts of the mind such that one part can count as deceiving the other. Another approach is to weaken the requirement that self-deception results in a *belief* and claim that it merely results in something like an "avowal" (roughly, a disposition to affirm something verbally); this way one avoids positing simultaneously the belief that *p* and the belief that ~p. Audi (1982, 1988), Rey (1988), and Funkhouser (2005) favor this approach. Still another approach is favored by Lazar (1999) and Mele (2001), according to which self-deception is not *intentional* and the contrary of the self-deceptive belief need not also be believed. Most of these investigations, however, while being interesting with respect to puzzles 1 and 2, do little to advance an understanding of *why* the capacity for self-deception exists as a property of the human mind.[5] Puzzle 3 is widely overlooked. One story sometimes told is that self-deception is the mind's means of protecting itself from psychological pain. But such an account does not solve puzzle 3, for misinformation, regardless of how content it might leave the organism, would seem inevitably to decrease fitness in the long run. *Pain*, physical and psychological, occurs for biological reasons.

Given that puzzle 3 remains unsolved, the time seems right for evolutionary psychology to enter the picture with an adaptationist theory of self-deception. After all, the initial conditions seem right: self-deception (as Wood points out in the quote above) is widespread; its prevalence may be taken to suggest that the capacity for it was a trait that was selected for in the ancestral environment. All that remains is for the evolutionary psychologist to posit a way in which self-deception could have increased fitness.

This strategy is precisely what Robert Trivers adopts in his article "The Elements of a Scientific Theory of Self-Deception."[6] On his account, self-deception serves at least two purposes that confer fitness benefits on the organism. First, the ability to self-deceive increases one's ability to lie effectively. Second, self-deception can serve the purpose of helping to orient one positively toward the future. If Trivers' theory is correct, we may take him to have solved puzzle 3. Further, if his work coheres conceptually with appropriate work in philosophy aimed at puzzles 1 and 2, we might take it to offer a definitive solution to the paradoxes of self-deception.

While Trivers' theory looks promising, I shall argue here that it fails for a number of reasons. Three criticisms are salient. First, the loose, qualitative predictions that the theory seems to have as consequences do not stand up to many kinds of cases of self-deception. Second, it makes implausible assumptions about what the phenotype set or variation would have been for natural selection to choose from. Third, Trivers' approach is most plausibly construed as a form what I will call *strong* adaptationism about self-deception: it makes the capacity for self-deception a modular feature of the mind. Such a view makes little sense of the connections that self-deception has to other aspects of human cognition. After leveling these criticisms, I argue that the capacity for self-deception can be much better accounted for as a *spandrel* in the sense that Gould and Lewontin (1979) give the term; it is a structural byproduct of other features of human cognitive architecture. After presenting the spandrel view, I discuss prospects for understanding the evolution of self-deception more generally. I distinguish a *weaker* version of adaptationism about self-deception than what Trivers endorses; the weaker form is not ruled out, although I remain skeptical. In any case, I maintain that any

successful theory of the evolution of self-deception will have to relate the capacity for it to its wider cognitive context.

**2. The Quarterback and the Cuckold: Conceptual Groundwork on Self-Deception**

As a starting point to discussion, let me first define *self-deception*.

Trivers characterizes self-deception as "active misrepresentation of reality to the conscious mind" (p. 114). This definition is inadequate. First, "misrepresentation of reality" suggests simply incorrect representation of facts.[7] But on this analysis *any* false belief that is consciously adhered to would count as self-deception, even ones that are clearly not self-deceptive (like the belief that the water cooler is in the hall, although someone has recently moved it to the lounge). Furthermore, this definition of self-deception *excludes* paradigm cases of the phenomenon, since a belief does not count as a "misrepresentation" if it turns out true. But take the case of the abused spouse who convinces herself that her husband will stop beating her for good after this time. I am inclined to count her mental state as self-deception, and to think so regardless of whether or not he (say) gets hit by a truck the next day and so does not beat her. But if he gets hit by a truck, the apparently self-deceptive belief will be true, and thus not count as "misrepresentation."

Rejecting Trivers' definition, I think we can best conceive of self-deception by considering two paradigm cases: the cuckold and the nervous quarterback. To start, imagine a man whose wife is cheating but who remains convinced that she's faithful. I am inclined to label this "self-deception" just in case the man in question has *compelling evidence* that she *is* cheating. She stays out all night, comes home with messed-up hair, is secretive about where she goes at night, etc. In fact, the self-deceived husband has

evidence such that, if he were to have that evidence about the behavior of another man's wife, he would conclude the woman was cheating. What makes the difference? The obvious suggestion is that he has a *desire* that his wife is faithful. This desire makes the causal difference in what he comes to believe. Now consider a nervous quarterback. He has strong reasons to believe his coach will be furious if he delivers a bad performance. Furthermore, he knows that this belief makes him nervous and that his resulting anxiety will cause him to perform badly. With this background he convinces himself that his coach will *not* be angry if he delivers a bad performance (for he *desires* this belief), despite *compelling evidence to the contrary*. To do this, he attends to what scanty evidence there is that the coach is a nice guy; if he succeeds in convincing himself, he's self-deceived.

These cases have two points in common. First, the self-deceived agent has epistemic access to compelling evidence to the contrary of the belief that he self-deceptively adopts. By "compelling evidence to the contrary" I mean evidence that under the usual application of the agent's epistemic norms would yield the belief *opposed* to the self-deceptive belief.[8] Second, in both cases there is a *desire* that is making the difference in what the agent believes.[9] Importantly, the existence of the desire confers *no further epistemic warrant* on the adopted belief. For, to take the first example, the agent's desire that his wife was faithful last night makes it no more likely that it is true that she was faithful. With these two points in mind, we arrive at the following definition of self-deception: *an agent is in a state of self-deception if and only if she holds a belief (i) that is contrary to what her epistemic norms in conjunction with available evidence would usually dictate and (ii) a desire for a certain state of affairs to obtain, or to have a*

*certain belief, causally makes the difference in what belief is held in an epistemically illegitimate fashion*.[10]

The difference between the quarterback and the cuckold is that the cuckold's desire that makes the difference is a desire for the world to be a certain way, whereas the quarterback desires a certain belief—it is not so much that he desires that his coach not be angry; mainly he desires that he *believe* that his coach will not be angry, since having this belief will calm his nerves. This analysis makes the cuckold's self-deception a species of wishful thinking, while the quarterback's might be called *intentional* self-deception. One might be tempted to call one self-deception and not the other, but I consider them both to be genuine cases of self-deception (I think general usage would lean in the direction of labeling both as cases of "self-deception").[11] In requiring that self-deception involve motivation at all, I believe am zeroing in on the same phenomenon on which other theorists studying "self-deception" have focused.[12] I would add, however, that on my view various emotions that have motivational effect and intentional content can also satisfy the second clause of my definition; for example, a fear that *p* may *play the role* of a desire that *~p* (without, of course, being reducible to a desire).

There may be types of motivation that figure into *other* forms of self-deception besides the ones I've identified so far. One possibility is that the agent, desiring that *~p*, may come self-deceptively to believe that *p*. For example, Othello, desiring that his wife *not* be cheating, came to self-deception in believing that she was. We may call this *dreadful self-deception*.[13] This case does in fact fit my definition, if we allow a certain flexibility in what we regard as an appropriate relationship between the content of the epistemic norm-subverting desire and the resulting belief. I won't discuss this form of

self-deception further, however, as I suspect it's underlying psychology is rather different from the two kinds discussed above (the quarterback and the cuckold); it also seems not to be the kind of self-deception that Trivers' has in mind.

Finally, I believe my definition conforms to the semantics of the term self-deception nicely for the following reasons. First, it stipulates that self-deception involves subversion of the evidence-based belief formation process (hence *deception*); second, it stipulates the subversion is due to one of the agent's *own* desires (hence *self*). These conceptual considerations, of course, do not defeat Trivers' project. His adaptationist explanation of the phenomenon might be correct, despite his bad definition.

**3. A Brief Overview of Trivers' Account**

Trivers offers essentially three hypotheses for why humans have the capacity for self-deception. First, self-deception may aid the deception of others, so it's adaptive. Second, self-deception may arise from internal representation of voices of significant others, where such "voices" can be ingrained genetically ("internal genetic conflict") or learned through contact. (This part of the theory I won't discuss, since whatever Trivers is talking about with it, it's clearly not self-deception, but something more like *being conflicted*; I include it in the summary for completeness.) Third, self-deception is adaptive insofar as it orients one favorably toward the future.

With respect to the first hypothesis, Trivers notes that hiding the fact that one is deceiving from one's own consciousness makes it less likely that one will inadvertently betray the fact that one is deceiving, since one does not realize it oneself. "Because deception is easily selected between individuals, it may also generate *self*-deception, the better to hide ongoing deception from detection by others" (p. 115). He then elaborates

on the hypothesis by giving a loose taxonomy of deceits that may be aided by self-deception: self-promotion, the construction of a biased social theory, and fictitious narratives of intention. Also noteworthy, Trivers posits *modules* to account for the self-deceptive hiding of true strategies from consciousness: "These [dishonest activities] can be thought of as directed by unconscious modules favored by selection so as to allow us to pursue surreptitiously strategies we would wish to deny to others" (p. 116).[14]

Turning to the second hypothesis, Trivers first focuses on parent-offspring interaction, noticing two points: (i) a parent has a 1/2 degree of relatedness to its offspring, whereas the offspring (naturally) has a degree of relatedness of 1 to itself; (ii) the instruction of parents can be enormously valuable to offspring. As a result, the offspring can benefit from parental instruction, although the interests of the parent and offspring are expected to diverge. From this Trivers concludes: " . . . it can easily be imagined that selection has accentuated the parental voice in the offspring to benefit the parent and that some conflict is expected within an individual between its own self-interest and the internal representation of its parents' view of its self-interest" (p. 122). The conflict of internal voices is then supposed to pave the way for self-deception.

Third, Trivers briefly considers the idea that positive illusions benefit humans. "Life is intrinsically future-oriented and mental operations that keep a positive future orientation at the forefront result in better future outcomes (though perhaps not as good as those projected)" (p. 126).

**4. Do Trivers' Views Stand Up?**

I will focus here on two aspects of Trivers' theory (and more the first than the second): his view that self-deception was selected for because it increases the ability to

deceive[15]; his view that self-deception was selected for because it orients one positively toward the future.

I take it as an assumption that having reliable information about the environment is beneficial for the survival and reproduction of any organism. I also take it as an assumption that this largely explains the cognitive mechanisms we have, such as vision, hearing, smell, touch, and taste. Faculties that provide a pathway for reliable information were selected for in the process of natural selection, not just in the more recent human environment, but also deeper in our evolutionary history. I doubt Trivers would disagree with these basic assumptions as a starting point for investigation.[16] Against this background, however, Trivers view is surprising. He is committed to the claim that a specific capacity for self-deception was selected for; such a capacity, if it did come about in this way, goes against our expectations about natural selection pushing us in the direction of having better information, because self-deception will generally (although not necessarily) tend in the direction of falsehood.

Trivers answer to this worry has to be that the ability to lie convincingly (hypothesis 1) has so great a fitness benefit that self-deception's contribution to it outweighs the loss of fitness we would expect to accompany something that detracts from reliable information flow. To start, one might doubt whether lying itself is really that beneficial in the first place, but this is not my criticism. It seems, rather, that self-deception would have to be highly targeted—that is, almost uniquely associated with situations in which lying is beneficial—in order to have enough fitness benefit to be selected for overall. To see this clearly, consider an early human who is asked whether there is food by the river. Perhaps this human is helped by being able to lie, because then

he has the food to himself. So given that there are many situations like this, self-deception, which will increase the ability to lie, will be selected for. But what would the self-deception here consist in? Well, convincing himself in the dialogue that there actually is no food by the river. But if he continues to believe this, then he ends up not getting the food at the river himself, *because he believes there is none*. In fact, the self-deceiver is worse off than if he had not lied at all, because then he at least would have gotten half the food.[17] Thus in order to confer the fitness benefit, the supposedly selected-for capacity for self-deception would have to produce self-deceptions that hold almost exclusively at the time when lying is useful. Even if such a cognitive feature were possible, it sounds very unlike the self-deception that we encounter in the real world.

Furthermore, although Trivers says nothing about how his theory might be tested, the considerations of the last paragraph suggest that there is at least one prediction that his theory makes that might facilitate testing of some sort. If Trivers were right, *then cases in which humans engage in self-deception could be expected to map closely to cases in which it is useful (potentially yields fitness benefits) to lie*. A systematic psychological study would be needed to test this, but I think there are two strong reasons for expecting this prediction will fail. First, it is perfectly possible (even frequent) for human self-deception to occur in cases where lying confers no benefit and is not even attempted. Second, human lying often occurs (and may even confer fitness benefits) without self-deception to help it along. To see the first point, consider again the quarterback from our earlier discussion who deceives himself into thinking that his coach will not be angry if he plays poorly. He may have no intention of spreading the belief to others; he engages in the self-deception for its effect on how he will play (better if less

nervous), not in order to lie. He may even hope that others think that the coach *will* be angry, because then they might pity him.[18] Such cases make it unlikely that the capacity for self-deception exists *because* it enhances the ability to lie. To see the second point, consider again the early human who lies about there being food by the river; here, presumably, there is no self-deception at all about where the food actually is. Putting the two points together, we see there must be plenty of self-deception without lying and plenty of lying without self-deception. So the main prediction that Trivers' theory would seem to make fails.

Another problem with Trivers' theory is that self-deceivers are often the only ones who believe what they're self-deceived about. Consider an average or below average driver who is self-deceived that he's a good driver. It's very likely that he's the *only* person who believes this. So it seems that his self-deception didn't help him convince anyone else. This kind of case is not merely hypothetical. Tara MacDonald and Michael Ross (1999) present a longitudinal study at the University of Waterloo that examines the accuracy of predictions student make about the future success of their romantic relationships. Comparing the lovers predictions to predictions of roommates and parents, the authors find that the people in the relationships themselves are significantly *less* accurate than the less optimistic others. If many of the people in the relationships were self-deceived about future prospects, which seems likely, then here we have a common type of self-deception in which *only* the self-deceivers themselves are taken in by the misinformation they propagate. We may take this to show that self-deception just does not do what it was selected to do in all cases. But it looks more and more like self-deception wasn't really selected to do what Trivers claims.

Trivers might respond to one of these points by saying that his point that self-deception yields a positive orientation toward the future explains self-deception that's unrelated to lying. But now it becomes increasingly unclear how Trivers' thinks this chapter in the history of natural selection is supposed to go. Is the capacity for self-deception supposed to be one trait or two—one to help lying and one to help positive orientation toward the future? If it is one capacity, what was the reason for its initial selection? If the initial selection was to aid lying, then my criticisms apply. If self-deception is supposed to come from two distinct modules (one for lying and the other for positive illusion), then the existence of one of them (self-deception to aid lying) should be doubted for the reasons given.

Thus it remains for me to discuss the possibility that self-deception, or a specific form of it, was selected to give humans a positive future orientation. Trivers writes:

> In the past twenty years an important literature has grown up which appears to demonstrate that there are intrinsic benefits to having a higher perceived ability to affect an outcome, a higher self-perception, and a more optimistic view of the future than facts would seem to justify. (p. 125)

Trivers cites the work of Shelley Taylor, who in various studies has demonstrated a correlation between having a "healthy" affect and having overly optimistic beliefs about one's situation. Conversely, depressed individuals are shown to have more accurate assessments of their situations.[19] The interpretation of such data advocated by Taylor (and Trivers following her) is that positive illusions "promote" things like "the ability to be happy"—where "promote" is meant causally. Let me now distinguish three things. First, there are the correlations that Taylor finds. Second, there's the interpretation of those correlations as showing that illusions *cause* mental well-being. Third, there's the view that self-deception was evolutionarily *selected* to produce the illusions that cause

mental well-being and hence positive future orientation. The first, the correlations, I have no reason to doubt. But the second, Taylor's interpretation, strikes me as misguided. Having inflated beliefs about one's abilities does not *cause* positive affect or positive future orientation. To see this, consider that many people who are *in fact* highly able and accomplished (and know this—thus they have positive beliefs) are not happy. And many with humble opinions of themselves do have a positive outlook. Furthermore, if someone were to burst a normally optimistic person's bubble and tell her that she's not a great driver and has only average looks, this would most likely induce temporary disappointment, but not a fundamental change in affect or future orientation. I would interpret the correlations that Taylor finds as indicative of causality in the opposite direction: positive affect causes positive illusions. Assessing Taylor's data thoroughly is beyond the scope of the paper, but it suffices here to make two points that follow from the preceding discussion: (i) the interpretation of her data is by no means settled; (ii) it is clearly psychologically *possible* for mental well-being and the kind of beliefs that result from positive illusions to exist independently. Self-deception is not *necessary* to avoid depression. Thus the burden is on Trivers to supply an argument for why self-deception would have been selected *as a means* to positive future orientation. None is given. In absence of such an argument, it is fair to look for another strategy for accounting for the existence of self-deception; I'll give one shortly.

Furthermore, if Trivers is to tell the adaptive story he wants to tell, he, like other adaptationists, has to assume that there was a high degree of variation in the phenotype set from which selection occurred. Otherwise nature will not have provided enough options for natural selection to sort through and pick out the things that ultimately are

adaptations. But here Trivers is in a delicate situation. If we assume too diverse a phenotype set, then it becomes clear that self-deception for the sake of lying will not have the optimal fitness value in comparison to the other possibilities. The reason is that one relevant possibility that we might put in the phenotype set assumed in our optimality model is the capacity to lie effectively *without* deceiving oneself. Then, because of the advantages of having reliable information over not having reliable information, this phenotype will win out over that of the self-deceptive liar (even if we suppose that self-deceptive lying is advantageous over not lying at all). If this is the case, then we do *not* expect self-deception that is tied to lying to win out. In order for Trivers account to work as an adaptationist story, he would have to assume a phenotype set from which selection occurred that did not include the possibility of non-self-deceptive liars competing with the hypothesized self-deceptive liars. Such an assumption would be highly puzzling, especially since there are many effective liars in existence who do not self-deceive. To put this objection more formally, Trivers faces a dilemma. Either there was a wide degree of cognitive phenotypic variation or not. If there was, then the ability to lie *without* self-deception would have won out; hence there would be no self-deception. If not, then self-deception that was targeted enough to aid other-deception in fitness enhancing ways is unlikely to have arisen.

We can extend an analogous criticism to Trivers' hypothesis that self-deception may have been selected also for its ability to orient one positively toward the future. Why would a self-deceptive positive future orientation win out over a self-honest positive future orientation? It seems likely that any benefits that arise from positive illusions could

just as well have existed as independent phenotypic characteristics. Against these, it is hard to see why the self-deceptive positive illusions would have won out.

There are many other problems that can be raised about Trivers' views on self-deception. One obvious one is the "grain" problem—the problem that evolutionary psychologists face in individuating what precisely were the problems set by nature that given adaptations arose to solve.[20] I wish to conclude my criticisms here by noting what on my view may be the greatest problem for Trivers' theory. He holds that the capacity for self-deception was a particular adaptation that was selected for and doesn't clearly relate it to other features of mind that might help make it possible. Thus he views self-deception as a sort of module (or module*s*) that arose in the history of human cognitive evolution. Presumably, then, this module would turn on whenever the adaptive benefits of deceiving oneself are at hand. Such a view is problematic insofar as it fails to relate self-deception to other features of cognition. Talbott (1995), for example, notes that biases of attention, memory, reasoning, and evidence gathering may all be involved in self-deception.

A view that makes the capacity for self-deception out to be a modular adaptation does not sit well with a view that makes it dependent on other features of human cognition. Although my arguments in this section have been directed against Trivers' modular account in particular, I think there are some general worries that cast doubt on the prospects for *any* modular account of self-deception. If self-deception arises from an adaptive module, then self-deception must facilitate some adaptive state of affairs X (where X is variable; for Trivers X would be "being a better deceiver"). But then the module posited must be quite complex: it must have means to (a) deceive the main belief

formation system, (b) detect when, for the purpose of bringing about X, it must do so, (c) have better information about the situation at hand than the main system (otherwise the misinformation it propagates won't count as deception), and (d) output situation-specific contents for variable situations. This is a tall order. In addition, reflection on cases shows self-deception can affect beliefs about a wide range of contents. People can be self-deceived about their hairlines, their intelligence, other people's intelligence, their favorite team's championship prospects, their own happiness, their intentions, the intentions of others, love, global warming, the competence of their local congressman, and many other subjects. A theory of self-deception as the product of a module for a specific purpose faces the challenge of explaining why a specific module should output such a variety of self-deceptions that are for the most part unconnected, many of which have no apparent adaptive purpose at all. These considerations are far from settling the issue, but I think they are enough to motivate looking for another style of explanation altogether.

In the next section, I develop a view that locates self-deception within a broader mental context. That is, I supply a (more or less) structuralist alternative to the adaptationist view. I should note that I appeal only to mental processes in developing my account whose existence is *independently* plausible. Thus my account will be more parsimonious by far than Trivers-style modular adaptationist accounts.

**5. Self-Deception as a Spandrel**

Gould and Lewontin (1979) begin their critique of adaptationism with the observation that any architectural structure that involves a dome mounted on top of rounded arches will have as a byproduct of this design what are called *spandrels*, tapered triangular surfaces that reside beneath the dome in the space between the arches. Gould

and Lewontin's point is that many phenotypic traits are analogous to spandrels; they are the result of the organism's structure, and it would be wrong to construe them as adaptations that were selected for in their own right, just as it would be wrong to construe the spandrels of a cathedral as spaces that the architect decided to include *independently* of the overall structural design. (The example often alluded to is the *chin*, which is the result of other structural features of the jaw, not a trait selected for in its own right.)

*Contra* Trivers, I think it is most fruitful to view the capacity for self-deception as a spandrel, not as an adaptation. That is, the human propensity for self-deception exists *not* because the trait was selected for in our evolutionary history, but because other aspects of our cognitive architecture have the capacity for self-deception as a byproduct.

To identify the features of mind implicated in the capacity for self-deception, let's first examine *how* a typical individual case of self-deception arises. Suppose Bob desires that his relationship with his girlfriend not end (*p*). Now this desire, being rather strong, causes Bob to feel discomfort when he notices signs that the relationship is ending. Given that such signs are abundant—for she is indeed planning to leave him—Bob has evidence that normally would lead him to believe that the relationship is over. But Bob, like most adult humans, is able to focus his attention selectively on some memories, perceptions, and beliefs at the exclusion of others. With this ability in the background, combined with a general inclination Bob has to avoid discomfort, Bob's attention focuses on what scanty evidence there is that the relationship is not ending and away from the abundant evidence that it is. But this scanty evidence, when focused on, leads Bob to believe what he wants to be the case: that his relationship with his girlfriend isn't ending.

Bob is in a mental state that satisfies my definition of self-deception. Bob believes contrary to the weight of his evidence, and the subversion of his epistemic norms is caused by a desire with content closely related to the content of the self-deceptive belief. In this case, the content of the relevant desire is the same as the content of the self-deceptive belief; thus I classify Bob's case as *wishful* self-deception, which is continuous with wishful thinking. Willful self-deception, as in the case of the nervous quarterback mentioned above, will involve similar selective attention, although the selective attention will be motivated more by explicit desires than general inclinations.

Let's abstract away some general features of mind from the analysis of Bob's self-deception. I hasten to add the qualification that my account probably is incomplete, since (i) other features may also be involved and (ii) I suspect these features by themselves are not *sufficient* to make complete sense of the phenomenon. I am neutral on whether or not the features I identify here themselves are adaptations, although I suspect some of them are.

The five features I have in mind are:

1. *A general inclination to avoid discomfort*. Humans generally find discomfort, including psychological discomfort, intrinsically aversive.[21]

2. *Selective attention to evidence*. Humans form beliefs in response to evidence that they find for those beliefs in the world. Often, evidence will uniquely determine what the beliefs are—if the wall appears white to me, I cannot do anything to make myself believe that it's purple. However, what evidence we attend to is influenced by what our interests are, and this will have an effect on what we come to believe.

3. *The inertia of the web of belief*. It is widely acknowledged in philosophy that beliefs do not typically occur without relation to a wide web of other beliefs that in some way justify them and give them content. I believe that one fact of human cognition is that the webs that constitute our belief sets typically have a degree of inertia; that is, a system of beliefs does not easily change entirely due to the existence of facts that are anomalous from the perspective of particular beliefs.

For the most part, I think this aspect of our cognitive economy is advantageous; for if our web of beliefs underwent revolution with each discovery of anomalous fact, we would be in a perpetual state of cognitive flux.

4. *Disposition to favor theories that are on the whole less complex*. The human mind (to use some terminology from computing) has limited storage space and limited computational power. Thus it is inclined to make sense of as much of reality as it can with the aid of as few informational resources as possible. This propensity may be seen as *informationally advantageous*.

5. *Awareness of self-fulfilling beliefs*. The holding of many beliefs contributes to the fulfillment of those beliefs. For example, I won't be able to jump across the ditch unless I believe that I can; but if I believe, I'm able. Thus it makes sense to form the belief. But this awareness that beliefs can be self-fulfilling can be misapplied. Believing I will win the lottery doesn't make it so.[22]

Before explaining formally how I think each of these features is implicated in self-deception, I should briefly review the empirical reasons to accept their existence, which are independent of any need to explain self-deception. If we regard anxiety as a form of psychological discomfort, we may take the research that LeDoux (1996) presents as favoring 1, the general inclination to avoid discomfort. LeDoux argues that the mammals he's studied find anxiety *intrinsically* aversive—they do not merely try to avoid the *cause* of the discomfort. Evidence for 2, selective attention—and the differential effect attention has on cognitive processing—dates back to the 1950s, when E. C. Cherry (1953) did an experiment in which subjects heard a different voice recording in each ear. Subjects in Cherry's experiment were asked to "shadow" (repeat the words of) only one channel; at the end of the experiment subjects showed no awareness of the contents of the non-shadowed channel. The inertia of the web of belief, 3, is supported by the ample literature on confirmation bias. Klayman and Ha (1987), for example, find evidence that people generally engage in what they call a *positive test strategy* when testing a proposition. "According to this strategy, you test a hypothesis by examining

instances in which the property or event is expected to occur (to see if it does occur), or by examining instances in which it is known to have occurred (to see if the hypothesized conditions prevail)" (p. 212). It is not hard to see how such a strategy will contribute to inertia in one's web of belief. Furthermore, social psychologists often discuss a *perseverance effect*,[23] by which they mean a general tendency not to change one's social beliefs even after those beliefs have been discredited. Most of the social psychology that deals with self-fulfilling beliefs (or "self-fulfilling prophecy") examines *interpersonal expectancy effects*. Rosenthal (1994), for example, reviews a long history of studies showing that teacher expectations of students influence the success of students in the direction of the expectations independently of whether or not the initial expectations had a basis in reality. I'm suggesting here that people have an awareness (usually implicit) of the self-fulfilling nature of many beliefs they have about themselves.[24]

Let me now state formally the contributions these five features make to the capacity for self-deception. We attend to evidence to varying degrees. This is generally beneficial, because if we were to attend equally to all evidence for any belief we might have, we would know a lot about many things that are of no interest to us and not enough about many things that are of much greater interest. But failure to attend to evidence that by our usual epistemic standards we should attend to, when that evidence is compelling and to the contrary of a belief we have, can leave us with beliefs that we ourselves ought not regard as justified. When that failure to attend to the evidence is due to a *desire* (or the discomfort that prospects of its non-satisfaction engender) that has no epistemic relevance to the truth of the belief in question, holding that belief becomes a case of self-deception.[25] That attention can be modulated by desires in a way that leads to self-

deception is important; for, as I claim in my definition, self-deception is constitutively caused by desires.

The inertia of the web of beliefs is a factor in self-deception as follows. The human mind is not easily disposed to giving up beliefs that are central to the web. Typically, much disturbance at the periphery is necessary. The fact that the web has inertia to begin with makes it possible to hang on to unjustified beliefs even when evidence to the contrary is compelling. Thus the inertia I allude to is a factor in the human capacity for self-deception—it contributes to the failure of epistemic norms and evidence that constitutes the irrationality of self-decpetion.[26]

Now suppose adopting a certain belief will add a great deal of complexity to one's belief set. A voter, for example, with incoming evidence that his favorite politician is corrupt in a particular way may have had prior reason to think that she is a good person, that she is honest, that he (the voter) is a good judge of character, that her past promises were made sincerely, etc. Against this background, acknowledging the politician's corruption would require the acceptance of a large number of additional facts. The politician is a mostly good person, *except* in certain particular ways; she is honest usually, but not in certain situations; he himself is sometimes a good judge of character, sometimes not; her past promises were sincere, but there were things about which she couldn't be completely forthcoming, etc. Of course, the voter might have to accept dubious explanations of certain news reports, but *not* acknowledging her corruption may reduce the informational complexity of his beliefs overall. We might generalize this and say that a greater deal of complexity in the web of beliefs *locally* can lead to less complexity overall. Of course, when the local complexity become too much, the

phenomenon of overall informational efficiency may start to contribute to the phenomenon of self-deception. Thus having informationally *advantageous* dispositions as part of our cognitive architecture contributes to our capacity for self-deception.[27]

Turning to 5, awareness of self-fulfilling beliefs, I note that many aspects of human performance require confidence. I must *believe* that I can do P in order to be able to do it. Thus many humans make a habit of pressuring themselves to believe they can do something as a normal means of trying to be able to. That is, they *desire* a belief of the form *I can accomplish P*. In cases where confidence actually affects performance, such belief formation can be justified, since the having of the belief that I can do something actually increases the likelihood that I really can. Such belief formation is easily extended to produce self-deception—particularly in cases where it would be manifestly irrational for me to believe (say) that I can do something very difficult. Or worse, it extends to cases where performance plays no role; people convince themselves they'll win the lottery "if I only believe."

Finally, here are some ways these features can work in groups. Features 4 and 2 can combine as follows: I don't want to increase the complexity of my belief set, so I ignore certain evidence. Feature 3 can combine with any of the other three, for example, with 5: if I used to be able to do something, then inertia of the web of beliefs combined with the thought that I only need confidence may help me self-deceptively hold on to the belief that I can. Feature 2 can also combine with 5: I ignore evidence that confidence makes no difference to a task.[28]

## 6. Better Prospects for Understanding the Evolution of Self-Deception

I should note before closing that the spandrel view of self-deception is not logically inconsistent with a *weaker* form of adaptationism than the modularist account that Trivers argues for. This form of adaptationism about self-deception would hold that the capacity for self-deception is indeed a structural byproduct, but one whose positive fitness value prevented it from being selected against (and hence extinguished). This is certainly a live possibility; it would, of course, require much more empirical evidence on the effects of self-deception to decide.[29]

But if other features of mind *are* implicated in the production of self-deception in the way that I have noted, then we should think in terms of selection of a *whole package* of features, with trade-offs being likely and abundant. The determinants of the fitness value of the package will be many, with self-deception being only one component. Whatever one's view on the fitness value of self-deception in particular, it's seems reasonable to think that being an extreme self-deceiver is maladaptive. However, if self-deception is a structural byproduct, then selection could not eliminate it entirely without eliminating the features of which it is a byproduct. But each of the features I identify appears very advantageous. A finite cognizer is greatly helped by being able to attend selectively to inputs and evidence, and will sometimes have to attend selectively to avoid mental discomfort for the sake of accomplishing another practical project. Likewise, the inertia of the web of beliefs can help creatures with limited processing time from being in perpetual cognitive flux. And, as noted, a disposition to favor theories that are on the whole less complex is informationally advantageous. Sometimes holding self-fulfilling beliefs is also useful—especially when the beliefs actually have good prospects for

coming out true. So even if there were selection pressure against self-deception, it is very hard to see which of the features that give rise to it *could* be given up without a severe blow being dealt to the normal functioning of the cognitive economy of the species in question—in this case, humans.

My own view, which must at present remain a conjecture, is that self-deception is not on the whole adaptive, but the features of mind that give rise to it are. Thus the capacity for self-deception comes as the downside trade-off of an overall package that is highly adaptive. If this view is plausible, then we at least have a candidate for a satisfactory solution to the puzzle of why self-deception exists (puzzle 3).

**7. Conclusion**

I have argued here that self-deception is not an adaptation of the sort Trivers advocates, but a spandrel, a byproduct of other features of cognitive architecture. One advantage of treating self-deception as a spandrel is that we can explain its continued existence despite the salient possibility that it is maladaptive. Self-deceptive conviction that someone is a friend can lead to being swindled, whereas one might not have been swindled if she were honest with herself. Self-deception about a partner's fidelity may lead one to put time and energy into raising a child that is not one's own. Self-deception can lead to the pursuit of goals that are a waste of time. The list goes on. Architectural plans often have undesirable byproducts (this, of course, is where the analogy breaks down, because *real* spandrels are often beautiful). I also wish to suggest that the qualitative predictions that my schematic theory seems to make are all quite plausible. First, if I am right about the role of the inertia of the web of belief, we would expect many cases of self-deception to be cases of maintaining old beliefs despite contrary

evidence that has piled up. For examples of this, consider Kuhn's (1962) accounts of paradigms that die only with the scientists who held them. Second, if my considerations about the web of belief and the role of favoring informational simplicity are correct, we would expect not that self-deception occurs regarding things that are close to the sensory periphery of the web, but rather that it would occur regarding more central beliefs that play a greater role in the simplicity of the overall system. This also seems plausible, for it is hard to imagine one deceiving oneself regarding whether one's socks are yellow, but self-deception often happens regarding central beliefs that have many connections to other practical beliefs about one's life (like whether or not a spouse is faithful).

To conclude, I wish to draw one general moral for the research program of evolutionary psychology. Trivers' attempts to explain self-deception in relative isolation from the many other features of the human mind to which it relates. Thus he finds himself in the position of attempting to explain why a feature of human psychology exists as though it were a feature in isolation, and the prevalence of this feature led him to consider it an adaptation. This sort of mistake is prompted by—and feeds back into—a picture of massive modularity of mind that evolutionary psychology often paints. The mind, it can be argued, does have modules, but it is not *all* modules. I believe that stepping back and considering the relations between features of the human mind (either from the perspective of psychology or philosophy of mind) *before* adverting to adaptationist stories will yield psychological theories that are both richer and more accurate.

**Acknowledgements**

For helpful exchanges on the subject matter of this paper, I'd like to thank: Alan Lloyd, Alfred Mele, John Perry, Erica Roeder, Elliott Sober, Kenneth Taylor, Robert Trivers, Avi Tuschman, the audience at the 2005 Columbia/NYU 5th Annual Graduate Conference in Philosophy, the audience at the Stanford Philosophy WIP Seminar, and two anonymous referees for *Philosophical Psychology*.

---

**Notes**

[1] Wood (1988).
[2] Jean-Paul Sartre (1956, p. 89) raises a closely analogous paradox regarding his concept *bad faith*.
[3] See especially Essays 9-11 in *Inquiries into Truth and Interpretation* and Essays 11 and 12 in *Essays on Actions and Events*.
[4] I am not so concerned to deal with puzzles 1 and 2 in this paper, although I believe my next section on the concept of self-deception offers a start. The literature, however, is extensive. For work relevant to puzzle 1, see (for example) McGlaughlin (1988) and Rorty (1988). Davidson's own work (1982, 1985, 1998) is largely addressed at puzzle 2. See Johnston (1988) for a view that takes puzzle 2 as a *reductio* of the interpretive view.
[5] Mele's work may be viewed as an exception to this claim. However, he is more concerned with how self-deception comes about rather than with the question of why the capacity for it exists.
[6] Trivers (2000). Trivers also adopts the strategy of explaining self-deception as selected to aid deception in Trivers (1985) and in the Forward to Dawkins (1976). I focus on Trivers (2000) because it is most recent.
[7] The first entry in the online OED for "misrepresentation" is: "Wrong or incorrect representation of facts, statements, the character of a person, etc."
[8] It may be further asked what I mean by "usual" here, since if "usual" just means non-self-deceptive then we are stuck with a circular definition. This problem, however, can be solved by saying that "usual" just means "in absence of the kind of belief-influencing desire that constitutes the second point of commonality between the two cases."
[9] Patten (2003) argues that a desire is not necessary for self-deception. On his view, an agent can become self-deceived about her own motives through mistakes caused by non-motivated biases of the sort that cause one to make mistakes about someone else's motives. Such mistakes about oneself are certainly possible, but it is questionable whether in their non-motivated form they should be viewed as self-deception. If there's no motivation involved in driving the epistemic failure, it might be more appropriate to call the resulting mistake an *honest* mistake.
[10] Of course, the content of the desire must be related to the content of the self-deceptively held belief, e.g., the cuckold in denial desires *that his wife is faithful* and this is what he self-deceptively believes, so in that case the relation of content is identity. But there are other possibilities besides identity. Nelkin (2002), for example, argues that the content of the motivating desire is *to believe* what ultimately comes to be self-deceptively believed. As my example of the quarterback indicates, I think this is one possibility. But how the contents of the two elements must be related is a subject of debate that wouldn't be an appropriate focus for this article.
[11] There is disagreement about which of these two notions is the right analysis of self-deception. McGlaughlin (1988) is explicit about the view that self-deception is on a continuum with wishful thinking, while Talbott (1995) adopts the "intentional" self-deception analysis. By defining the term in the inclusive way that I do, I avoid the controversy. Of course, I have independent reasons for adopting the analysis I do.

[12] For a discussion arguing that some sort of motivation is needed for a case of irrationality to be a case of self-deception see Mele (2001, pp. 104-110), who argues against Martha Knight's (1988) purely cognitive explanation of certain forms of self-deception. In my view, it may be that other types of irrationality that are not motivated have etiologies that are similar to the etiology of what I call self-deception. That is, they will be spandrels of *some* of the same cognitive features of which the capacity for self-deception is a spandrel. This would make them "cousins" of self-deception without being precisely the same phenomenon as that in which most philosophers and psychologists who have discussed self-deception have been interested. Those who build motivation into their constitutive characterization of self-deception include: Gur and Sackeim (1979), Bach (1981), Audi (1982, 1988), Davidson (1982, 1985, 1998), Pears (1984), Rey (1988), Mele (2001), Nelkin (2002), Funkhouser (2005), and others. This paper is unfortunately not the place to argue conclusively that motivation must be included in the definition of self-deception.

[13] Mele (2001) calls this "twisted self-deception."

[14] The grammatical antecedent of "These" in this quote is actually somewhat unclear, since there is no plausible plural noun in the sentence prior to the one quoted. Trivers does, however, seem to be describing a general kind of deceptive activity that self-deception may aid, and "These" may refer to various activities falling under that kind. It is also not perfectly clear that Trivers is positing a self-deception module or modules, but his emphasis on modules on pages 116-117 make this the most likely interpretation.

[15] Depending on how one understands "deceive," if one is truly self-deceived in having a certain belief, then propagating that belief to others will not count as "deceiving," because one has the belief oneself. But I trust it will be clear what is meant in this context: "deceiving" is standing in for "conveying beliefs that one would not under normal circumstances believe oneself." In this text, I often use "lie" instead of "deceive" for stylistic reasons.

[16] In presenting this argument to others, I've encountered the objection more than once that skittish animals like mice are counterexamples to the claim that having good information channels is generally adaptive. The objection is that a mouse is better adapted for believing any hovering shadow to be a hawk, even when it's not true that it is. The first thing to note is that, if this really is a counterexample, it's an exception to a rule that holds widely. The mouse's eyes, ears, nose, and whiskers send its brain information that on the whole is processed in a highly sophisticated and reliable way. In any case, the tendency towards overrepresentation of *hawk* in the mouse's cognitive processing has an easy explanation. This is not so for self-deception, which can infect a very large variety of beliefs—ranging from beliefs about lovers to sports to politics—and often has destructive consequences. In any case, it is clear that Trivers has the same back ground assumption in mind when asking the following rhetorical question: "In trying to deal effectively with a complex, changing world, where is the benefit in misrepresenting reality to oneself?" (p. 115) In other words, Trivers is in agreement with me in taking the subversion of internal information that goes on in self-deception as something to be explained. I'm arguing here that his explanation isn't adequate.

[17] Ramachandran (1998, pp. 278-279) has an example in his endnote 3 on Trivers' view that parallels my own. Ramachandran's example is of a chimp deceiving another chimp about the whereabouts of bananas, who, if he's self-deceived, won't get the bananas either. Ramachandran says this seems to him to be an "internal contradiction" in Trivers' theory, although he does go on to speculate how it might be resolved.

[18] I note that it would be a weak objection to this point to say that it does not hold because there were not quarterbacks in the ancestral environment; whatever that environment was, it will not be that difficult to construct analogous cases (assuming the environment was social and people were at all disposed to get nervous for social reasons).

[19] See, for example, Taylor and Brown (1989) and Taylor (1998). I'd like to thank one anonymous reviewer for encouraging me to address this issue.

[20] For a discussion of this see Sterelny and Griffiths (1999, p. 352)

[21] This is a general rule that has many exceptions—forms of behavior that seem aimed at obtaining discomfort, e.g., masochism. Nevertheless, the general rule holds widely and suffices for present explanatory purposes.

[22] These five features of human cognition are, of course, not at all new ideas. What I think is original, if anything, is the way I employ 3 and 4 to understand the phenomenon of self-deception (2 and 5 have figured in other discussions) and the insight that the way such features work together to produce self-deception affects what sort of biological stance we may take on the phenomenon. My inspiration for 3

comes from Quine (1953) and Kuhn (1962). I realized that 4 may play a role in self-deception on reading Sober (1981, pp. 103 ff.). Feature 2 is discussed in Talbott (1995).

[23] See Aronson *et al*. (2005, p. 67) for a discussion of the perseverance effect.

[24] The one feature I cite for which I was unable to find support in the form of experimental data is 4, the disposition to favor theories that are on the whole less complex. My thought on the matter is, however, motivated by the following kind of case. Suppose you're a detective faced with one suspect who has a simple explanation for being near the scene of the crime and one with a complicated explanation. I think there's a natural inclination to give the simpler explanation more credit.

[25] I note here that this analysis leads to a possible connection between self-deception and deception in which the direction of explanation is the opposite of what Trivers gives (by this I mean that attempts to deceive can explain some cases of self-deception). Often in order to deceive, one must gather evidence in support of the deception. The biased accumulation of evidence can then lead the deceiver himself to be deceived.

[26] See the first clause of my definition.

[27] The "inclination" to favor less complex theories on the whole should not be read as satisfying the desire component of my definition of self-deception. The inclination is rather a cognitive tendency that can contribute to the violation of epistemic norms in self-deception. This means of course that feature 4 cannot by itself cause a phenomenon satisfying the definition of self-deception to come about, so other features, e.g., 1, 2 and/or 5, may be required.

[28] Of course, some of these features could combine to produce spandrel byproducts that do not count as self-deception. See endnote 11.

[29] I'd like to thank one anonymous reviewer for making me aware of the possibility of weaker forms of adaptationism about self-deception.

## References


Aronson, E., T. D. Wilson, and R. M. Akert (2005) *Social Psychology*, Upper Saddle River, NJ, Prentice Hall.

Audi, R. (1982) "Self-Deception, Action and Will," *Erkenntnis* **18**(2): 133-158.

Audi, R. (1988) "Self-Deception, Rationalization, and Reasons for Acting," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press**:** 92-120.

Bach, K. (1981) "An Analysis of Self-Deception," *Philosophy and Phenomenological Research* **41**(3): 351-371.

Cherry, E. C. (1953) "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America* **25**(5): 975-979.

Davidson, D. (1980) *Essays on Actions and Events*, New York, Oxford University Press.

Davidson, D. (1982) "Paradoxes of irrationality," in *Philosophical Essays on Freud*, edited by R. Wollheim and J. Hopkins, Cambridge, Cambridge University Press.

Davidson, D. (1984) *Inquiries into Truth and Interpretation*, New York, Oxford University Press.

Davidson, D. (1985) "Deception and Division," in *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, edited by E. LePore and B. P. McLaughlin, Oxford, Blackwell.

Davidson, D. (1998) "Who is fooled?" in *Self-deception and Paradoxes of Rationality*, edited by J.-P. Dupuy, Stanford, CSLI**:** 1-18.

Dawkins, R. (1976) *The Selfish Gene*, Oxford, Oxford University Press.

Funkhouser, E. (2005) "Do the Self-Deceived Get What They Want?" *Pacific Philosophical Quarterly* **86**(3): 295-312.

Gould, S. J., and R. Lewontin (1979) "The Spandrels of San Marco and the Panglossian Paradigm," *Proceedings Of The Royal Society of London, Series B* **205**: 581-598.

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

Johnston, M. (1988) "Self-Deception and the Nature of Mind," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press**:** 63-91.

Klayman, J., and Y.-W. Ha (1987) "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review* **94**(211-228).

Knight, M. (1988) "Cognitive and Motivational Bases of Self-Deception:Commentary on Mele's Irrationality," *Philosophical Psychology* **1**: 179-88.

Kuhn, T. S. (1962) *The Structure of Scientific Revolutions*, Chicago, University of Chicago Press.

Lazar, A. (1999) "Deceiving Oneself or Self-Deceived," *Mind* **108**: 263 - 290.

LeDoux, J. (1996) *The Emotional Brain*, New York, Touchstone.

Lockie, R. (2003) "Depth psychology and self-deception," *Philosophical Psychology* **16**(1): 127-148.

MacDonald, T., and M. Ross (1999) "Assessing the accuracy of predictions about dating relationships: How and why do lovers' predictions differ from those made by observers?" *Personality and Social Psychology Bulletin* **25**(11): 1417-29.

McLaughlin, B. P. (1988) "Exploring the Possibility of Self-Deception in Belief," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press.

Mele, A. R. (2001) *Self-Deception Unmasked*, Princeton, Princeton University Press.

Nelkin, D. K. (2002) "Self-Deception, Motivation, and the Desire to Believe," *Pacific Philosophical Quarterly* **83**: 384-406.

Patten, D. (2003) "How do we deceive ourselves?" *Philosophical Psychology* **16**(2): 229-246.

Pears, D. (1984) *Motivated Irrationality*, Oxford, Oxford University Press.

Quine, W. V. O. (1953) "Two Dogmas of Empiricism," in *From a Logical Point of View*, edited Cambridge, Harvard University Press.

Ramachandran, V. S., and S. Blakeslee [coauthor] (1998) *Phantoms in the Brain*, William Morrow & Company.

Rey, G. (1988) "Toward a Computational Account of *Akrasia* and Self-Deception," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press**:** 264-296.

Rorty, A. O. (1988) "The Deceptive Self," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press**:** 11-28.

Rosenthal, R. (1994) "Interpersonal Expectancy Effects: A 30-Year Perspective," *Current Directions in Psychological Science* **3**(6): 176-179.

Sartre, J.-P. (1956) *Being and Nothingness*, H. E. Barnes (tr.), New York, Washington Square Press.

Sober, E. (1981) "The Evolution of Rationality," *Synthese* **46**: 95-120.

Sterelny, K., and P. E. Griffiths (1999) *Sex and Death*, Chicago, University of Chicago Press.

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

Taylor, S. E. (1989) *Positive Illusions: Creative self-deception and the healthy mind.*, New York, Basic Books.

Taylor, S. E. (1998) "Positive Illusions," in *Encyclopedia of Mental Health*, edited by H. S. Friedman, San Diego, CA, Academic Press. **3:** 199-208.

Trivers, R. (1985) *Social Evolution*, Menlo Park, CA, Benjamin/Cummings.

Trivers, R. (2000) "The Elements of a Scientific Theory of Self-Deception," *Annals of the New York Academy of Sciences* **907**: 114-131.

Wood, A. W. (1988) "Self-Deception and Bad Faith," in *Perspectives on Self-Deception*, edited by A. O. Rorty and B. P. McLaughlin, Berkeley, University of California Press**:** 207-227.