Editorial Manager(tm) for Philosophical Studies
Manuscript Draft

Manuscript Number: PHIL1441R1

Title: Finite Rational Self-Deceivers

Article Type: Original Research

Section/Category:

Keywords: self-deception; rationality; rational; philosophy of mind; psychology; paradox; evolution; evolutionary pscyhology; spandrel; adaptation; adaptationism; finite; finitude; human nature; belief; belief formation; epistemic norm; byproduct; irrational; irrationality; attention; selective attention; web of beliefs; cognitive dissonance; interpretive view; Donald Davidson; epistemic; epistemology

Corresponding Author: D. S. Neil Van Leeuwen,

Corresponding Author's Institution: Stanford University

First Author: D. S. Neil Van Leeuwen

Order of Authors: D. S. Neil Van Leeuwen

Manuscript Region of Origin:

Abstract: Abstract: I raise three puzzles concerning self-deception: (i) a conceptual paradox, (ii) a dilemma about how to understand human cognitive evolution, and (iii) a tension between the fact of self-deception and Davidson's interpretive view. I advance solutions to the first two and lay a groundwork for addressing the third. The capacity for self-deception, I argue, is a spandrel, in Gould's and Lewontin's sense, of other mental traits, i.e., a structural byproduct. The irony is that the mental traits of which self-deception is a spandrel/byproduct are themselves rational.

Finite Rational Self-Deceivers

D. S. Neil Van Leeuwen
Stanford University, Department of Philosophy

Abstract: I raise three puzzles concerning self-deception: (i) a conceptual paradox, (ii) a dilemma about how to understand human cognitive evolution, and (iii) a tension between the fact of self-deception and Davidson's interpretive view. I advance solutions to the first two and lay a groundwork for addressing the third. The capacity for self-deception, I argue, is a *spandrel*, in Gould's and Lewontin's sense, of other mental traits, i.e., a structural byproduct. The irony is that the mental traits of which self-deception is a spandrel/byproduct are themselves rational.

Introduction: a Paradox, a Dilemma, and a Tension

Three puzzles plague the notion of self-deception.

First, the classic *paradox* of self-deception points to an apparent incoherence in the very concept. If A deceives B, then A does not believe what she makes B believe. So if A deceives A, A must not believe what A believes. Thus the existence of self-deception seems to entail a contradiction. But manifold examples of human behavior are appropriately called cases of self-deception. So how is self-deception possible?[1]

Second, knowledge is *prima facie* critical to the evolutionary success humans have had. Our vision, hearing, other senses, and higher cognitive abilities largely seem to owe their existence to the fitness value of knowledge. The widespread human capacity for self-deception, however, undermines knowledge. The prominence of self-deception, therefore, creates a *dilemma* about how to understand the fitness value of knowledge in general. If knowledge enhances fitness, then the capacity for self-deception shouldn't exist since it undermines knowledge. But it does exist. If knowledge does *not* enhance fitness, then the existence of self-deception is not a problem for evolutionary theory. But we are then deprived of the most obvious explanation of our most complicated and interesting traits. How shall we solve this dilemma?

The third puzzle is motivated by the work of Donald Davidson, who indeed recognized it: the widely held interpretive view of the mental holds that rationality is constitutive of the mental, including beliefs.[2] One cannot sensibly attribute beliefs to an agent unless those beliefs make rational sense to the person attributing them. But self-deception is an irrational belief state that people, with apparent justification, attribute often. Thus there is a *tension* between holding that rationality is constitutive of belief and holding that belief can come about by self-deception. How shall we resolve this tension between the force of the interpretive view and the fact of self-deception?[3]

In this paper, I advance solutions to the first two puzzles. I also, by explaining how self-deception relates to rational capacities and evidence, lay a groundwork for addressing the third. I do this by advocating three theses.

1. I offer a new definition of self-deception [Thesis 1], one that captures the epistemic tension inherent in saying one deceives oneself, while avoiding the absurdity of saying one believes what one doesn't. I state the definition fully in section 1.

2. I advance *the byproduct view* of self-deception [Thesis 2], which runs as follows: *The capacity for self-deception is a byproduct of a number of critical abilities humans have that enable us to cognize and behave rationally given finite minds*. Self-deception, in short, is a byproduct of finite rationality.

The byproduct view takes a synchronic perspective on the capacity for self-deception, but it also makes salient a diachronic explanation of that capacity.

3. I advance the view [Thesis 3] that the capacity for self-deception is what biologists Steven J. Gould and Richard Lewontin call a *spandrel*. The thesis, precisely, is: *The capacity for self-deception is a structural byproduct of features that were not selected for their role in the production of that capacity*. I do not claim here that the features I cite are themselves adaptations; it is enough for my view that there is independent justification for positing their existence apart from their role in self-deception.

Theses 2 and 3 are identical as to the features they claim the capacity for self-deception is a byproduct of, but they differ as to the claims they make *about* those features. Thesis 2 claims that they are critical to finite rational cognition and behavior. Thesis 3 claims that they did not arise in natural selection *for* producing self-deception.

The definition is meant to solve the paradox. Theses 2 and 3, the byproduct and spandrel theses, solve the second puzzle in a way that does not force one to deny the fitness value of knowledge. As a spandrel, the historical existence of the capacity for self-deception rode piggy-back on the evolution of those features of which it is a spandrel. Since those features are rational, we can explain the capacity for self-deception as the downside of a tradeoff: selection favored a package of features that are themselves rational, but have an irrational byproduct. The whole package will then have to have been more fit than competing packages, but it will be the rationality of it that contributes to the fitness—not the self-deception.

It should be apparent that the byproduct view is of independent philosophical interest. Rationality is of course a contested concept, but for clarity I offer the following working definition: a capacity is rational if it is (a) conducive to truth and coherence in an agent's belief set or (b) conducive to means-end coherence in practical planning, and (c) is not directly detrimental to either (a) or (b)[4]. The cluster of features I identify constitute one way of having rationality in a finite mind with practical projects, but that cluster yields the capacity for self-deception as a byproduct. Is there some other way of having rationality in a finite mind? My discussion will suggest that, for any finite creature with human-like desires, rationality brings with it the capacity for self-deception.

I structure this paper as follows. In section 1, I present my definition of self-deception and explain how it resolves the paradox. In section 2, I defend my theses about the *capacity* for self-deception, arguing for both theses in parallel. I then describe, in section 3, a paradigm case of self-deception and apply my framework to it. To explore the wider implications my theory has, I argue in section 4 that the spandrel view of self-deception challenges the dominant adaptationist paradigm in evolutionary psychology. Section 5 concludes with implications about self-deception and human rationality.

1 Defining Self-Deception and Resolving the Paradox

On the way to defining self-deception, I start with three paradigm cases.

Imagine a college dropout who's aware of several important pieces of evidence that suggest finishing his degree will improve his job prospects. He's seen statistics on earnings; his brother who graduated has fared much better at getting jobs; and several positions he wants require a degree. But he wants it to be the case that his chances are good even without finishing; he becomes self-deceived that they are.

Now imagine a soccer player who's made nervous by the possibility that her coach will be angry if she plays poorly. She might think: "I wish I believed coach wouldn't be mad. Then I wouldn't be nervous and could play better." The weight of evidence suggests coach is the angry type, but by focusing on the scanty evidence that coach is nice our soccer player deceives herself into believing the coach won't be angry.

As a last example, recall Othello, who deceives himself into believing that Desdemona *isn't* faithful. He wants her to be faithful and has evidence that largely suggests she is. But he self-deceptively believes she's *un*faithful; disaster ensues.

Now let me present some theory-neutral terminology that will allow us to discuss aspects of self-deception. First, *the product of self-deception* is the mental state that results from self-deception and in some sense accepts the content that the self-deception is *about*; I hold the product is a belief. Second, the *deceptive element* is the state or attitude in one's own mind that subverts the normal belief formation processes. I hold this is a form of motivation and will use the word "desire" for most of my discussion as a general way of referring to this motivational component.[5] Third, the *doxastic alternative* is the proposition that the agent in some sense *should* believe. The doxastic alternative for the college dropout is: *finishing my degree is needed to improve my job prospects*.

If we return to the paradox, we find two major classes of views aimed at defining self-deception in a way that resolves it. First, there's the class of views that holds the product of self-deception is a belief; when I'm self-deceived that *p*, I believe that *p*. The paradox is then resolved either by drawing a division in the mind between the deceived part (which believes *p*) and deceiving part (which believes ~*p*), or by denying the requirement that the self-deceiver also believes ~*p*.[6] The second class of views holds that the product of self-deception is an avowal or avowed belief, a mental state underlying verbal behavior, but lacking important properties of genuine beliefs, such as deep connections to actions aside from merely verbal ones. The belief view is more common, and is held by, among others, Davidson (1998), Lazar (1999), Mele (2001), Pears (1984), Talbott (1995), and myself. Variants of the avowal view are held by Audi (1988), Rey (1988), and Funkhouser (2005).

The avowal view seems to provide an easier solution to the paradox: the self-deceiver doesn't both *believe* that *p* and believe that *~p*, since the way in which *p* is held is merely an avowal.

The problem with the avowal view is that it resolves the paradox by denying deep connections between self-deception and action. But consider a gambler who marches to the casino self-deceived that he can climb his way out of debt; he's *acting* on his self-deception. I hold the belief view, because the self-deceived gambler wants to get out of debt and *believes* that he can do so by gambling. Likewise, Othello's self-deception explains not merely his verbal behavior, but also his *action* of killing his wife. The product of self-deception plays the same role in action explanation as other beliefs, so it is best viewed as a belief.[7]

Let us then hold fixed that the product of self-deception is a belief, and see what our paradigm cases have in common.

First, the agents aren't forming the self-deceptive belief out of lack of intelligence or relevant evidence; the belief goes *contrary* to their epistemic norms (rational rules for belief formation) and the evidence they have[8]. If the soccer player had similar evidence about someone *else*'s coach, she would correctly conclude *that* coach was the angry type. Second, a *desire*, with content related to the belief formed, is making the causal difference in the belief formation process. That the desire makes the causal difference will be clear from a thought experiment; suppose Othello just didn't care whether Desdemona was faithful. Then, I think, his self-deceptive belief wouldn't have arisen. On the basis of these considerations, I give this definition [Thesis 1]:

> *An agent is in a state of self-deception if and only if*
> *(i) she holds a belief,*
> *(ii) that belief is contrary to what her epistemic norms in conjunction with what evidence she has would usually dictate, and*
> *(iii) a desire, with content appropriately related to the belief formed, causally makes the difference to what belief is held in an epistemically illegitimate fashion.*[9]

"Epistemically illegitimate" here is to be understood as *illegitimate relative to the epistemic norms that the agent actually has*, since believing contrary to norms one doesn't have shouldn't count as self-deception (although perhaps it would be wishful thinking).[10] A self-deceived agent, I hold, does not just violate standards of rationality; she violates *her own standards* of rationality.[11] My definition is meant to resolve the paradox of self-deception as follows: it captures the epistemic tension[12] involved in self-deception by requiring that the agent have compelling evidence to the contrary of the belief formed, but by not requiring both beliefs be held it escapes positing the psychological absurdity (if it is absurd) of holding two contradictory beliefs.

What relation must obtain between the content of the deceptive element (the desire) and the product of self-deception (the belief)? There are three possibilities and thus three main types of self-deception. First, the content of the desire can be the *same* as the content of the resulting belief; I call this **wishful self-deception**, or self-deception that's continuous with wishful thinking. The college dropout, for example, has wishful self-deception. He believes what he wants to be the case. This kind of self-deception isn't planned; one slides into it. Second, the desire can be to have *the belief* that's ultimately formed, as with the nervous soccer player; I call this **willful self-deception**. Pascal, it seems, was willful. Finally, the content of the desire can be contrary to the content of the resulting belief, as in Othello; I call this **dreadful self-deception**[13].[14]

The chart below summarizes forms of self-deception and related phenomena. $\mathbf{D_i}\boldsymbol{p}$ means i desires that *p*; $\mathbf{E_i}\boldsymbol{p}$ means i has evidence that by her epistemic norms on the whole favors believing *p*. These are all different forms of irrationality agent i can have in believing that *p* ($\mathbf{B_i}\boldsymbol{p}$). The far left column refers to the deceptive element.

	Non-self-deceptive belief	**Self-deceptive belief**
	$\sim\mathbf{E_i}\boldsymbol{p}, \sim\mathbf{E_i}\sim\boldsymbol{p}, \mathbf{B_i}\boldsymbol{p}$	$\mathbf{E_i}\sim\boldsymbol{p}, \mathbf{B_i}\boldsymbol{p}$
$\mathbf{D_i}\boldsymbol{p}$	wishful thinking	*wishful self-deception*
$\mathbf{D_i}\mathbf{B_i}\boldsymbol{p}$	willful thinking	willful self-deception
$\mathbf{D_i}\sim\boldsymbol{p}$	dreadful thinking	dreadful self-deception

The idea behind the wishful thinking, willful thinking, and dreadful thinking column is that the agent, with desire playing a causal role, forms a belief without having compelling evidence that it's true or compelling evidence it's false. The various forms of self-deception are stronger forms of irrationality, for in those cases the agent has evidence that supports the belief that runs *contrary* to the belief that's formed—in fact, evidence that supports the negation of what's believed. We can see, then, that there is a continuum between wishful thinking and self-deception, with the line being crossed when the evidence contrary to the product is sufficient by the agent's epistemic norms to support the doxastic alternative.

In what follows, I assume my definition and conceptual framework are plausible and address what the capacity for self-deception consists in and why it exists, focusing on *wishful self-deception*, which seems the most common form.[15]

2 The Etiology of Self-Deception

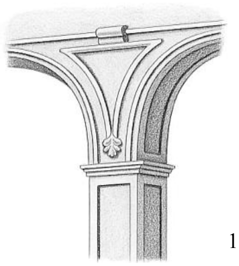
The second puzzle of self-deception, the dilemma about the fitness value of knowledge, raises the question of why the capacity for self-deception exists.

Prima facie, the capacity for self-deception looks bad for fitness, so it's puzzling that it hasn't been selected out. Many examples of self-destructive behavior caused by self-deception support the intuition that it has negative fitness value: people get taken advantage of in relationships they self-deceptively believe are good, get in fights they can't win but self-deceptively believe they can, fail to fix serious problems they're self-deceived about, and so on. Nevertheless, one approach to this puzzle is to argue that self-deception is not really of negative fitness value—or is even of positive fitness value. This is the approach that Trivers (2000) takes in arguing that self-deception enhances our ability to lie. Note, however, that this thesis relies on the controversial empirical hypothesis that the capacity for self-deception enhances fitness, or at least did in the ancestral environment.

My approach, represented in the spandrel thesis, will have two major advantages over the other. First, insofar as the capacity for self-deception arises from features of mind whose existence is independently plausible, my solution is more parsimonious for not having to posit additional functions. We won't have to posit anything besides what we already know exists; the work comes in showing *how* these features give rise to self-deception. Second, my approach does not rely on a dubious empirical postulate: that self-deception enhances fitness.

Let me now develop the spandrel view.

Gould and Lewontin (1979) begin their critique of adaptationism in biology with the observation that any architectural structure with 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.


[16]

Their point is that many phenotypic traits are analogous to spandrels; they are the result of the organism's structure; 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. They give the example of the *chin*, which results from other structural and developmental features of the jaw and wasn't selected for its fitness value.

I hold that the capacity for self-deception is an evolutionary spandrel. This thesis may be broken into two components. First, the capacity for self-deception is a structural byproduct of other features of mind. Second, the existence of those features in evolutionary history is most plausibly construed as due to functions *other* than their contribution to self-deception. To argue the first, I demonstrate in the following sections what mental features the capacity for self-deception is a by product of. To argue the second, I develop—concurrently with my explanation of how those features contribute to self-deception—*the byproduct view*: the features in question contribute in key ways to the ability we have as finite creatures, insofar as we have that ability, to cognize and behave rationally.

Let me flesh out the outline of this argument in slightly more detail. My central argument for the spandrel thesis will be to define what I think is the *essential complex* of features of mind out of which self-deception—wishful self-deception in particular—

arises. I show that the operation of the features in this complex to produce self-deception is wholly intelligible without appeal to any mechanism whose "purpose" (evolutionarily or otherwise) is to produce self-deception. I then highlight other features that aid the process of self-deception; these can be divided into epistemic *facilitators*, factors that in certain contexts make it easier to believe contrary to usual epistemic norms, and *desire sources*, aspects of mind that give rise to the sorts of desires that easily figure in self-deception. Throughout I will make the case that the features identified are conducive to rational cognition, goal-setting, and goal attainment given finite limits. If this is true, then the capacity for self-deception is not merely a quirk that arises out of unimportant features of what our brains are like; it's rather a capacity that arises from features essential to human rationality.[17]

2.1 The Essential Complex

Seven features of the human mind comprise the essential complex that yields self-deception. The first three are:

EC1: *Desires have a characteristic sting that accompanies anticipation or evidence of their non-satisfaction.*[18]

EC2: *Humans have the ability to attend selectively to inputs and evidence.*

EC3: *Humans have a general inclination to avoid discomfort.*[19]

How do these features of mind contribute to self-deception? Say I desire that *p*. Thus I also have the disposition to feel a sting when ~*p* is anticipated. Because of this desire that *p*, when I encounter evidence of ~*p*, I have a feeling of discomfort or queasiness (EC1). This evidence must be cognized *as evidence* of ~*p* in order for it to cause the discomfort—otherwise it simply wouldn't bother me. This discomfort is there especially when my attention is on the ~*p* evidence, but abates when my attention shifts—especially when it shifts (EC2) to *p* evidence (what little there may be).[20] Against a background inclination to avoid discomfort (EC3), my attention shifts to the *p* evidence. Self-deception ensues, for the focus of attention on the *p* evidence gives rise to the *p* belief, even though the *total* evidence possessed in conjunction with epistemic norms would dictate believing *~p*.

Thus runs the basic causal chain from the desire that *p* to the belief that *p*, despite compelling evidence that *~p*. Importantly, it isn't the case that the normal functions[21] of EC1-EC3 are to produce self-deception. EC1, the sting of desires, is conducive to goal-attainment; it prods us out of situations in which our desires are unlikely to be satisfied. EC2, selective attention, is an essential feature of any finite cognitive system with interests; without it the mind would be mired in a wash of inputs, most of which would be irrelevant to the organism's ends. EC3, general inclination to avoid discomfort, typically has the function of keeping one out of situations that may be harmful and is thus generally conducive to goal attainment. The irony of EC3 is that, by helping make self-deception possible, it sometimes contributes to keeping one *in* harmful situations about which one is self-deceived.[22]

The essential complex also includes:

EC4: *Evidence in the human mind is structurally organized.*

We wouldn't be able to attend to evidence selectively if that evidence weren't organized in ways that allow us to search through it. The structure of the Evidence Box, so to speak, is essential to our ability to cognize rationally, for without it selective attention would be aimless. But it also is an essential component in our capacity for self-deception. Thus, to relate this point explicitly to Thesis 2, we have the surprising result that self-deception is, at least in part, a byproduct of *rational* capacities. (By way of contrast, an *infinite* rational being wouldn't need evidence in her mind to be structurally organized, for infinite powers would allow attending to *everything* at once, with no need to search through the evidence.)

The essential complex, as I've described it so far, presupposes the following feature:

EC5: *Humans form beliefs on the basis of evidence in conjunction with epistemic norms.*

The deep irony is that this rational feature of mind is both subverted and implicated in the same self-deception. It's subverted with respect to total evidence, but implicated in the formation of the self-deceptive belief by the evidence selectively attended to.[23]

For completeness, I add:

EC6: *Humans experience pleasure at evidence that their desires will be satisfied.*

EC7: *Humans seek pleasure.*

An example of EC6 in action is the exuberance felt at the end of a sports competition when victory is in sight. It should be clear immediately how EC6 and EC7 contribute to attainment of the goals that one has, but they are also implicated in the comfort-driven modulation of attention that produces self-deception.[24]

2.2 Facilitators and Desire Sources

A *facilitator* is a mental feature that makes easier the failure of epistemic norms and evidence in self-deception. A *desire source* gives rise to desires that are apt to cause self-deception. Not just any cause of desire will do, because desires whose content concerns subject matters it's hard to be self-deceived about for epistemic reasons will not typically give rise to self-deception. Here I present and explain (first) the most important facilitators and (second) the most important desire sources in the capacity for self-deception.

F1: *The web of belief has inertia.*

Beliefs typically occur in relation to a web[25] of other beliefs that justify them and give them content. One fact of human cognition is that the webs that constitute our belief sets have inertia. A system of beliefs does not easily change entirely due to the existence of facts that are anomalous from the perspective of particular beliefs. This aspect of our cognitive economy is largely advantageous to coherence in our belief sets; for if our webs of beliefs underwent revolution with each discovery of anomalous fact, we would be in a perpetual state of cognitive flux. It would not make sense to perform a massive *modus tollens* on our entire web simply because one belief encounters anomaly. This explains to some extent why confirmation bias, when it involves the tendency to look for confirmation of things already believed, may be helpful; it helps shield us from constant cognitive revolution.

The dark side of the inertia of the web is that it facilitates self-deception. First, it makes it easier to hang on to beliefs under the influence of a desire even when the evidence available has become compelling in the other direction. The essential complex may not always be sufficient by itself to cause a cuckold to hang on to the belief that his wife is faithful against the weight of the evidence. But given inertia of his already existing web of beliefs, self-deception becomes possible. Kuhn (1962) discusses scientists who won’t adopt a new paradigm even when the rest of the scientific community has signed on; some of these are probably cases of inertia-aided self-deception. Furthermore, the inertia of the web helps explain self-deception in the teeth of strong evidence to the contrary of the self-deceptive belief: multiple self-deceptions in a context can underlie the conscious reclassification of evidence and holding of beliefs; in such cases the inertia of the web is the mental difference between a house of cards and a house of cards with glue.

F2: *Desires and other emotions generate conjectures and thoughts.*

Antonio Damasio, in *Descartes' Error* (1994), argues that two of the brain's emotional centers, the ventro-medial pre-frontal cortex and the amygdala, are critical for reasoning. He doesn't differentiate between theoretical reason, reasoning to form true beliefs, and practical reasoning, reasoning to form good actions; his thesis is more plausible when taken to be about practical reasoning. But one of his points about the role of emotions in life planning suggests how emotions can influence theoretical reasoning as well. Emotions, including desires, cause ideas to be suggested in our mind for consideration. When one desires approval, for example, thoughts of oneself being approved of and accepted constantly pass through consciousness. Emotions themselves cause the appearance of representations in thought ("images," as Damasio puts it) of the objects at which the emotions are directed. This, of course, facilitates planning and practical reasoning about life choices, and Damasio shows that patients with damaged emotional centers are incapacitated to act rationally: due to lack of appropriate emotions the right ideas often don't occur to them for consideration. *But insofar as emotions and desires propose thoughts, conjectures, and hypotheses for the agent to consider, they will also influence the formation of beliefs—theoretical reasoning*. This is simply because we can't form beliefs about contents that haven't occurred to us. This mode of influence of desires and emotions on belief formation is beneficial for both cognitive and practical goal-attainment, for without the guidance of emotion directing what we attend to our cognitive capacities would be mired in consideration of an infinitude of useless information.

We often have desires for things intangible but felt to be important, such as what will happen in the future, the abilities of our children, the attitudes of friends and enemies, and events we won't ever see. In short, we have desires for things about which the beliefs fall far from the sensory periphery of the web of beliefs. When such desires that *p* arise, the *p*-thoughts arise also, as explained. The perpetual recurrence, however, of *p*-thoughts due to desires that *p* can be the first step in the self-deceptive formation of the belief that *p*; for that same desire can, in virtue of the essential complex, cause selective attention to evidence in favor of the belief that *p*. Since the *p*-thought is the first step in the formation of the *p*-belief, the property the desire has of proposing representations for thought, and of occupying the mind with them, thus makes it easier for self-deception to occur.

F3: *Humans can apply differing degrees of skepticism to different propositions.*

Skepticism—the withholding of belief from a proposition that does not meet a certain level of justification—comes in different degrees, ranging from refusing to believe a newspaper account without hearing it corroborated elsewhere to deep Cartesian skepticism about the external world. Most humans can modulate the degree of skepticism applied in given circumstances; citizens of a democracy, for example, will typically apply a higher degree of skepticism to information from a party they oppose than to information from the party they support. Being skeptical is a rational capacity, for it helps screen out belief in falsehoods. But applying skepticism differentially to different propositions can make us less truth-tracking, not more, whenever the differential application of skepticism is caused by desires other than the desire for truth. Discomfort with evidence in favor of believing that *~p*, caused by a desire that *p*, can make a person want to reclassify that evidence; differential application of skepticism can aid in reclassifying evidence and thus in the self-deception that *p*.

F4: *Humans can suppress unwanted memories.*

The debate over whether there is suppression of unwanted memories, which seems to face a paradox similar to the one facing self-deception, has gone on for nearly a century. It has, however, recently been put to rest by Anderson *et al.* (2004) who produced not only evidence that there is such suppression, but also fMRI evidence concerning the neurobiology underlying it.

> Memory suppression requires people to override or stop the retrieval process. Lateral prefrontal cortex is involved in stopping prepotent motor responses, switching task sets, and overcoming interference in a range of cognitive tasks. It may be hypothesized, therefore, that people suppress unwanted memories by recruiting lateral prefrontal cortex to disengage hippocampal processing. (p. 232)

Suppression is not tantamount to self-deception, since it can have the useful independent function of minimizing distraction when we must handle a situation to which the suppressed memory does not pertain. Memory suppression can thus facilitate rational thought, whenever the item suppressed needs to be bracketed in order for the agent to reason properly. But memory suppression can facilitate self-deception. Suppression may be used deliberately in willful self-deception to undermine memory of evidence contrary to the desired belief. But it may also figure into wishful self-deception without there being any intention to self-deceive. Quite simply, the discomfort engendered by certain memories can make us wish to suppress them, which we do. When that discomfort is caused by a desire that *p*, the suppression of memories that provide evidence that *~p* can facilitate self-deception by lowering the epistemic bar for the *p*-belief.

I now turn from epistemic facilitators to *desire sources*.

DS1: *Humans form complicated intentions.*

According to Michael Bratman (1987), an intention plays at least three roles in mental life. First, it causes one to engage in *means-end reasoning* about how to accomplish the intended goal. Second, it causes one to *screen out options* that would undermine the intended goal. Third, it causes one to *track* the intended goal. I hold that intentions, in virtue of playing these roles, engender what I'll call subsidiary desires—desires for states of affairs that would be conducive to the realization of the intended goal. Tracking engenders desires for information on the intended target; means-end reasoning brings about desires for realization of means to the end; screening creates desires to be rid of obstacles.

Why are subsidiary desires to intentions apt to give rise to self-deception? Often they're not, as when an intention to travel to Chicago involves a subsidiary desire to get to the airport; whether or not I've arrived at the airport on time is not something it's easy to be self-deceived about. But often, especially in the case of long-term intentions, subsidiary desires of intentions will be for things not immediately tangible but still felt to be important. If a parent intends to send a child to a good college, this intention will be accompanied by a subsidiary desire for the child to do well when he gets to high school—this is a means to an end. But suppose this subsidiary desire arises when the child is badly underperforming in eighth grade. In this sort of situation, the subsidiary desire concerning the child's intelligence may be the sort that gives rise to self-deception, since evidence about a person's intelligence is often mixed enough, even if it's heavily weighted to one side, to enable the kind of selective attending involved in the essential complex.

Do the other two functions of intentions, tracking and screening, also generate desires likely to be implicated in self-deception?

It's less likely that the tracking function frequently brings about desires ripe for self-deception. The main purpose of tracking is accuracy and knowledge, so its subsidiary desires won't be for states of affairs as much as information. Even if the goal of the intention is not well-defined, such as sending a child to a "prestigious" university, the tracking of related facts, such as admissions standards and tuition costs, will focus on concrete things about which self-deception occurs only rarely[26]. I also don't see that the screening function of intentions is a major source of desires that produce self-deception, since it will mainly be the source of desires *not* to do particular things. Nonetheless, the means-end reasoning function of intentions on its own is sufficient for intentions to be an important desire source for self-deception.

DS2: *Humans experience cognitive dissonance when their behavior diverges from the norms of their self-concepts.*

Cognitive dissonance is the discomfort people feel when their behaviors don't conform to their conceptions of themselves as moral, competent, and consistent. As an aversive drive, we want to be rid of it; thus we *want* circumstances in the world to be such that our behavior doesn't turn out to be immoral, incompetent, or inconsistent after all. Such desiring may well be beneficial to accomplishing practical projects, since it may motivate changing incompetent, inconsistent, or immoral behaviors. But such desiring is also conducive to self-deception, for it involves contents not near to the sensory periphery of the web of beliefs but nonetheless felt to be important. In Aronson *et al.* (2005), Elliot

Aronson recalls an otherwise affable Vietnam veteran he met who believed the Vietnamese weren't human. Clearly, the balance of evidence he has is on the side of the belief that they *are* human; this explains why he felt guilty at first about the atrocities he committed. But because of dissonance he desired that it not be the case that the beings he killed really were human; this desire is the deceptive element in his self-deception that they weren't. To us, it's plain that the civilians he killed were human; it would have been plain to him too had it not been for self-deception. But the point is that the question of whether or not a certain entity counts as *human* is sufficiently metaphysical and sufficiently debatable, as we've seen from debates on abortion, that it falls into the sphere of possible self-deception. Cases like that of the veteran can be multiplied.

*   *   *

Before moving to my discussion of a paradigm case, let me re-cap the argument thus far. The challenge was to explain why humans have the capacity for self-deception. The steps I took to meeting this challenge were as follows.[27] (1) I explained what features of mind are involved in the capacity for self-deception. (2) I argued—concurrently with saying what they are—that those features are rational. The byproduct view follows: (3) the capacity for self-deception is a byproduct of rational features. Once we've come this far, two arguments will get us to favoring the evolutionary view that the capacity for self-deception is a spandrel over the view that it is an adaptation. (4) Holding it is a spandrel is perfectly consistent with the *prima facie* plausible view that knowledge enhances fitness. (5) Holding that it is a spandrel is more parsimonious, for we don't need to posit an additional adaptive function for self-deception.

3 Application of the Spandrel View to a Paradigm Case

My focus here has been to explain why and how wishful self-deception in particular exists (as opposed to willful or dreadful self-deception[28]). The answer to the question of why we have the capacity for self-deception is that the capacity is a byproduct of mental features, each with useful functions, which in combination yield self-deception. To see how the capacity works, let's consider a paradigm case.

The case of wishful self-deception that I gave in section 1 was that of the college dropout. The case I discuss here is more tragic, but also common. My focus is abuse in familial or spousal relationships. Understanding the psychological causes of the behavior of the abuser is utterly important, but we should also ask about the psychology of the abused. Does self-deception ever play a role in keeping a person in a relationship in which he or she is abused? I know personally of two such cases; I don't think they're unusual. In one case, a woman abused by her boyfriend admitted, after the relationship ended, that she was self-deceived that the abuse would end. In another case, a man explained, years into middle age, that he had been self-deceived up until age 25 that his mother was a good parent, despite an ongoing pattern of abuse when he was a child. There is no reason to doubt these people's accounts. How does self-deception arise in such cases? I'll present my understanding of the case of the abused romantic partner.

This person has many cognitions—experience of abuse, knowledge of patterns of abuse—that on her own epistemic norms *should* yield the belief that her boyfriend will continue to hurt her if she stays in the relationship. But she has an *intention* (DS1) to have a healthy relationship; this intention has the subsidiary desire that the abuse stop even with the relationship continuing. This desire is the deceptive element. Although she has compelling evidence for the belief that the abuse will continue, she also has some limited evidence (in the form of repeated promises from the boyfriend) that it will stop.

Because of her desire for the abuse to stop and the relationship to be healthy, she feels the sting of disappointment (EC1) when attending to evidence suggesting abuse will continue and comfort when considering memories of promises that suggest it won't (EC2 and EC6). The general inclinations to avoid discomfort and find pleasure (EC3 and EC7) cause her attention (EC2) to be directed to the comforting evidence. Attention to this evidence, by EC5, completes the self-deception. She believes the abuse will stop even with the relationship continuing.

This process is aided by the inertia of the web of belief (F1), because she started out the relationship believing there would be no abuse and having many other beliefs about the goodness of her boyfriend supporting this belief. Some of her memories of abuse, furthermore, become suppressed (F4). She greets her friends who tell her to get out of the relationship with a higher degree of skepticism (F3) than she usually has. Because of her desire for the relationship to work, thoughts of relational harmony repeatedly occur to her (F2), making it easier for her to hold to the product of self-deception. The cognitive dissonance (DS2) arising from the tension between her conception of herself as a self-respecting person and her behavior of staying in an abusive relationship adds to her desire that the abuse stop in a way that further facilitates self-deception.

This is how, by a spandrel of the mind, an intelligent woman with ample evidence of her boyfriend's abusive character becomes self-deceived. There is no need to posit an intention to self-deceive or an adaptation that's designed for self-deception.

4 Self-Deception, Spandrels, and Methodology in Evolutionary Psychology

I hold that the theory of self-deception presented here has significant implications for the field of evolutionary psychology.

It's no secret that the dominant paradigm in evolutionary psychology is adaptationist. Consider this passage from Tooby and Cosmides (1995):

> Our cognitive architecture resembles a confederation of hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestors. Each of these devices has its own agenda and imposes its own exotic organization on different fragments of the world. (p. xiv)

If this is representative, then the program of evolutionary psychology has been to explain basic mental phenomena as features that were advantageous in the ancestral environment. Within this paradigm, Robert Trivers (2000) attempts to explain self-deception as an adaptation to help other-deception. I criticize Trivers' theory extensively elsewhere[29] on grounds of internal difficulties. But my main criticism of the dominant paradigm in evolutionary psychology at present—which also applies to Trivers—is that there is not enough serious consideration of alternative evolutionary hypotheses to adaptation. There are some examples of byproduct views within evolutionary psychology; art, fiction, religion, and music have all been claimed to be byproducts.[30] (Note that *byproduct* by itself is a wider category than *spandrel*, since something can be a byproduct of phenotypic traits combined with cultural facts. "Spandrel," on my understanding, denotes a byproduct of phenotypic traits.) But those byproduct claims tend to be about cultural phenomena restricted to specific content domains and to not basic, highly general mental processes like self-deception. The structural complexity of the human mind should prompt us to consider the possibility that many more prominent features of mind than have been recognized so far are structural byproducts—spandrels.[31] I hope that this essay not only explains the capacity for self-deception in particular as a spandrel, but also provides an example of how one might construct a spandrel theory in general.

There are many critics of evolutionary psychology. But there is a fallacious implicit assumption that many of them seem to share with those they criticize: either a prominent mental tendency is an adaptation or it is the result of environmental influence. This is no doubt simplifying the dialectic, but one gets the impression all traits are either like teeth (adaptations) or like hairstyles (environmental). The battle is then fought over which traits fall into which category, with evolutionary psychologists advocating adaptations, their critics advocating environment, and moderates advocating a "complex interplay" of the two. In this dialectic, spandrels are often ignored. When the possibility of spandrels does come up, rarely is a systematic theory developed of how the spandrel in question actually works.[32]

I suspect the problem might be that no compelling general approach to constructing psychological spandrel theories has been exemplified. In an Appendix on evolutionary psychology that summarizes the views of evolutionary psychologists Robert Wright and Richard Wrangham, a RAND Monograph Report presents Wright's defense of evolutionary psychology against some spandrel-type criticisms that have been made by Steven Jay Gould: "Wright responds that, while the criticism must be well-taken, Gould has never given an example of such a 'spandrel' in the literature" (p. 69). Lacking a good example of a mental spandrel, it seems, many evolutionary psychologists have felt free to pursue the adaptationist paradigm unencumbered. Furthermore, a close examination of this Appendix reveals an important and curious presupposition of the authors. Gould's arguments about spandrels are presented as criticisms of "evolutionary psychology," without appending any qualifier to the phrase "evolutionary psychology" such as "adaptationist." This shows that the authors implicitly equate evolutionary psychology with adaptationist evolutionary psychology. I think there is no good reason for this. Understanding the evolution of the mind/brain should involve both spandrel and adaptationist explanations—and probably many other kinds as well.

Now an example of a mental spandrel has been given; I have not only claimed that self-deception is a spandrel, but also shown what features self-deception is a consequence of and how self-deception follows from them. If the methodology I recommend is taken seriously, this could change the landscape of evolutionary psychology significantly. My methodological recommendation is this: in seeking to explain the etiology of a certain mental phenomenon, consider *first* in detail what other features of mind (or body) that mental phenomenon may depend upon; take seriously the possibility that the phenomenon in question is merely a byproduct of those other features without having been selected for in its own right; only *after* such consideration should one attempt to reconstruct an evolutionary history. This recommendation really amounts to this: understand the *anatomy* of a phenomenon before attempting to reconstruct its *evolutionary history*. Doing so will yield a picture of mind that is less modular, richer, and more accurate. Ironically, following these recommendations could lead to better and more developed adaptationist theories as well, for the interconnectedness of various adaptations would become much more apparent.

5 Self-Deception and Human Nature

What is the human mind such that it deceives itself? Or, what are the fundamental properties central to the capacity for this mental state? My answer is this: we have capacities for rational logico-mathematical and evaluative judgment, but we need to cognize and survive with finite powers in the limits of finite time. If we had infinite rational processing that allowed total objective formation of judgments of good and bad and true and false, then we wouldn't self-deceive. In the theoretical ideal of a perfectly rational angel, such judgments form the basis for all actions. In animals, finite perception, memory, and processing necessitate a much more immediate cause of action in the context of hostile environments: desires.

Desires alone, however, don't give rise to self-deception. *Self-deception comes from rationality in the context of a finite desiring mind*. As I've stressed, many of the features of which self-deception is a byproduct are rational. The ability to attend selectively to evidence is rational; it allows us to sift through the wash of inputs and attend to the relevant ones. The formation of beliefs on the basis of evidence is rational. The inertia of the web of beliefs is rational; it prevents perpetual cognitive flux and may be seen as characteristic of finite rational cognition as such. The ability to apply differing degrees of skepticism is rational; it can be used to help us screen out falsehood. And the ability to form complex intentions is rational; it facilitates theoretical and practical achievements. But all of these rational abilities are implicated in the pervasive form of human irrationality called self-deception.

Thus the capacity for self-deception is not an incidental addition to our cognitive make-up. It comes from rationality in the context of finitude, two features thought central to being human since the ancient Greeks. To summarize, our finitude contributes to our capacity for self-deception in two ways: (i) it necessitates *desires* for the sake of survival;

(ii) it necessitates several forms *selectivity in cognitive processing*. When the desires (i) influence the selectivity (ii), one is on the road to self-deception.

Humans are finite rational self-deceivers. Does this mean self-deception is inevitable for humans? I don't know either way, but I do think the *propensity* to self-deceive is inevitable. Self-deception can, however, be better avoided by cultivating cognitive habits that neutralize the aspects of mind that give rise to self-deception. One can confront discomforting evidence and accept it for what it is. I have also argued that there are kinds of desires that humans characteristically have that are more likely to engender self-deception: moral desires, desires for dispositional states of mind to obtain, and desires for intangibles that are felt to be important. One can be aware of these. A person with such desires and mixed evidence is in a context ripe for self-deception.[33]

Endnotes

[1] There are actually different formulations of the paradox. The one I give here presents self-deception as apparently involving a straightforward contradiction: belief and not belief. But, on a weaker formulation, attributing self-deception involves attributing a belief that *p* and a belief that ~*p*, which seems to involve positing something psychologically absurd without directly resulting in logical contradiction on the part of the attributer. For either formulation the strategy for resolution will be the same: give a definition that captures the phenomenon without positing something absurd.

[2] See Davidson (1982, 1985, 1998).

[3] There are not only these three puzzles. These are the most visible tip of the philosophical iceberg of problems surrounding self-deception. Self-deception raises questions about: agency, moral responsibility, self-knowledge, the role of emotion in cognition, the distinction between what is perceived and what is inferred, what beliefs are in general, doxastic voluntarism, the distinction between tacit and explicit beliefs, belief attribution, and the value of truth.

[4] It's important to note that this is a working definition. I'm sure possible situations could be imagined in which this definition is satisfied by an irrational capacity and a deviant causal chain, but this should do for now. Also, the word "directly" in clause (c) is important, since, as we'll see, rational capacities are indirectly—via self-deception—implicated in decrease in truth in the belief set.

[5] Why can't other kinds of emotional pollution besides desires muddle the epistemic process in a way that counts as self-deception? Reflection suggests that self-deception may be caused by fear, jealousy, love, hate, shame, and even values and ideology. My response in defense of my definition will not be to exclude these, but to include them insofar as they have a motivational aspect. An aimless fear or paranoia cannot be the deceptive element in self-deception, although they can cause other cognitive foibles. But love or fear can be if they are directed in certain ways and motivates.

[6] See note 1.

[7] I've recently become aware of empirical data that supports the belief view. Ramachandran (1998) discusses an experiment he does in which a patient in denial about her left side paralysis grasps a tray on one side with her right hand instead of in the middle, suggesting she *believes* that she could use her left hand to pick up the other side. I also argue against the avowal view in other work.

[8] When I say that an agent has evidence that *p*, I mean that she has cognitions that by her own epistemic norms constitute evidence that favors believing *p*.

[9] There are several other definitions of self-deception on offer in the philosophical literature. The one closest to my own is Mele's (2001). See also Robert Audi (1988), Kent Bach (1981), and George Graham (1986), to name a few.

[10] This naturally raises the question of what it means to have an epistemic norm. There are many complications lurking that do not fit into the scope of this essay. In general, however, I think that self-deceivers count as self-deceived because they believe contrary to patterns of belief formation (i) that they typically follow and (ii) that are rationally justifiable. In other work, I define a very weak form of self-deception in which a person believes contrary to norms that she does hold but that are not rational in the first place. I suspect intuitions will diverge as to whether such a state is appropriately labeled "self-deception." I avoid these complications here, however, and focus on paradigm cases.

[11] The sentence to which this is an endnote raises an important point. One cannot simply eliminate the mystery of the existence of self-deception by saying that the human mind is not that rational; for in self-deception one violates epistemic norms that one actually *has*. Violating one's own norms, I hold, is more mysterious than violating abstract cannons. That's why the relativization of self-deception to the agent's norms is so important. As far as I know, I am the first to define self-deception in a way that makes use of this kind of relativization.

[12] See Audi (1997), who argues that capturing such a tension is critical to a good definition of self-deception. Audi criticizes Mele for not requiring anything that entails the tension. My definition, I hold, is not susceptible to such criticism.

[13] Mele (2001) refers to it as "twisted self-deception." In labeling this "dreadful," I mean to suggest that the person engaged in this type of thinking or self-deception is motivated by some form of dread or fear. I am not suggesting that these types of irrationality are, from an outside perspective, in any way worse than the other forms.

[14] It may be that there are more kinds of self-deception than the three discussed here that one can categorize on the basis of relation between content of motivational element and content of belief. It is my impression that most cases of self-deception will be classifiable as one of these three types. But the "appropriately" in the content clause of my definition is meant to leave open the possibility of other types

[15] Lazar (1999) and Mele (2001) agree that non-intentional forms of self-deception are the most common.

[16] Image: *The American Heritage Dictionary of the English Language*: Fourth Edition, 2000.

[17] Alfred Mele (1997, 2001) also has put forth developed views on the etiology of self-deception. In other work, I compare our theories in detail, but a detailed comparison would be out of place in this essay. So I'll confine myself to pointing out the two most important differences. First, Mele does not address anything resembling the byproduct or spandrel theses, so my work speaks more directly to the second two puzzles identified in the Introduction. In fact, he doesn't raise these two puzzles at all. Second, Mele's FTL model will have difficulty accounting for cases of self-deception that the spandrel model handles well, i.e., cases in which the self-deceived agent comes to have a belief, the holding of which can only be seen as having high subjective costs.

[18] Different desires have their own characteristic stings; for some the sting is annoyance, for others anxiety, and for still others simple disappointment. I take the features I identify here to be normal aspects of the human mind: they obtain for the vast majority of adult people.

[19] LeDoux (1996) makes the point in the cases of fear and anxiety, for example, that it's not just the anxiety-causing stimulus that people seek to avoid, but the anxiety itself.

[20] The fact that evidence needs to be mixed in order for self-deception to happen is crucial for explaining a very important feature of self-deception. Self-deception very rarely occurs concerning perceptual beliefs; in occurs much more often concerning what might be called intangibles. The reason for this is that evidence for or against perceptual beliefs is mixed far less often.

[21] When I use the word "function," I am using it in the pre-theoretic sense without a commitment to the functions being adaptations (although, as indicated, I think they might be).

[22] The question naturally arises *why* the modulation of attention should be subject to the general inclination to avoid discomfort. One answer is that psychological discomfort of any sort is very often distracting in a way that prevents accomplishment of goals, so one may need to attend away from what causes discomfort to accomplish something else.

[23] This point is also important for addressing the third puzzle. One might argue that the fact of self-deception is a *reductio* of Davidson's interpretive view. (Cf. Johnston (1988, p. 88): "Rational connections are not constitutive and exhaustive of the mental.") But this inference would be too quick, for the involvement of EC5 in self-deception shows that there is a rational step in this (overall) irrational phenomenon.

[24] Throughout this discussion, I've been talking about desires as the form of motivation that gives rise to self-deception. One might reasonably wonder to what extent my account will generalize to cases in which emotions like anger, love, or jealousy constitute the motivational element behind the self-deception. I hold there are three relevant possibilities. First, desires are often associated with emotions and caused by them, so that in such cases the explanation I can give will simply have the added clause that an emotion was the source of the deceptive element desire. For example, if I am jealous of a rival, I may want him to be unworthy; this wanting will then be the deceptive element in self-deception that he is unworthy. Second, often intense emotions can cloud judgment without intervening desires, causing one to have unjustified beliefs. Some might loosely call this self-deception, but it is far enough from paradigm cases of self-deception that such talk strikes me as misleading. The phenomenon of simply having worse judgment when one is angry, for example, is no doubt interesting, but it is not appropriately labeled self-deception. Third, however, it is at least a conceptual possibility that an emotion should constitute the motivational/deceptive element in self-deception *without* and associated desire doing the causal work that I posit here. The important question, then, will be: to what extent does my account generalize to *those* cases? Suppose, for example, Harry's *disgust* with Jennifer for stealing in the past leads him to believe self-deceptively that her children are morally blameworthy also. Matters are tricky here, for this sounds like a somewhat borderline case of self-deception. But I do think that important aspects of my account will generalize. Selective attention will still, in this case, be driven by finding some evidence comfortable to attend to and other evidence not, and what causes the difference in what's comfortable will still be the motivational state, i.e., the emotion. Furthermore, the facilitators I discuss in following sections, like the inertia of the web of belief, can all still come into play. I will have to leave a fuller discussion of this topic, however, to another occasion.

[25] Of the sort Quine (1953) talks about. Ramachandran (1998, ch. 7) discusses clinical cases of denial in ways that suggest the sort of view I'm advocating.

[26] This comment raises the following difficult philosophical issue: what is the difference between perceptual beliefs and beliefs arrived at via inference? This paper is not the place to address this topic. But it seems to be that self-deception does not typically occur for perceptual beliefs—although see Ramachandran (1998, ch. 7) for some extreme cases.

[27] These steps are not meant to represent the order of presentation. Rather this is one way of representing the logical structure of the argument.

[28] Extending this model to cases of willful and dreadful self-deception will involve examining how the kinds of desire constituitively involved in *those* types can trigger the sorts of selective attention and other processes here identified. For reasons of space, I have not explored such an extension in this paper.

[29] Van Leeuwen (forthcoming).

[30] See, for example, Bloom (2005) and Pinker (1997).

[31] Paul Bloom has pointed out to me in email that, just as a matter of logic, the adaptations there are (however many this happens to be) will have as a consequence many more byproducts; furthermore, most mainstream evolutionary psychologists are well aware of this. While this is true, the point doesn't exempt the dominant paradigm from the criticisms I offer here; it is one thing to observe that there must be spandrels and another thing to work out rigorous theories of how they might work. Furthermore, the centrality of the mental processes hypothesized to be spandrels is also being raised here; my point is that central mental processes like self-deception that greatly and generally influence behavior have been largely exempt from the kind of theorizing offered in this paper.

[32] I am also not suggesting that "adaptation or environment or spandrel" is an exhaustive disjunction; there is drift and other factors that may influence phenotype as well.

[33] Work on this paper has benefited from exchanges with Kenneth Taylor, John Perry, Krista Lawlor, Dagfinn Føllesdal, Erica Roeder, John Gabrieli, Elliott Sober, Joanna Fidducia, Lanier Anderson, Michael Friedman, Nadeem Hussain, Helen Longino, Marc Pauly, Joshua Landy, Al Mele, Robin Jeshion, Jesse Prinz, Ron Suny, Hans Ulrich Gumbrecht, Paul Bloom, and Robert Trivers.

References

Anderson, M. C., K. N. Ochsner, B. Kuhl, J. Cooper, E. Robertson, S. W. Gabrieli, G. H. Glover, and J. D. E. Gabrieli (2004) "Neural Systems Underlying the Suppression of Unwanted Memories," *Science* **303**: 232-235.

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

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.

Audi, R. (1997) "Comments," see Mele (1997).

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

Baron-Cohen, S. (1995) *Mindblindness: An Essay on Autism and Theory of Mind*, Cambridge, MA, MIT Press.

Bloom, P. (2005) "Is God an accident?" *Atlantic Monthly* **296**(5): 105-112.

Bratman, M. E. (1987) *Intentions, Plans, and Practical Reason*, Cambridge, Harvard University Press.

Damasio, A. (1994) *Descartes' Error*, New York, Avon.

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

Fukuyama, F., and C. S. Wagner (2000). Information and Biological Revolutions: Global Governance Challenges—Summary of a Study Group. *RAND Monograph Report*, RAND Corporation.

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.

Graham, G. (1986) "Russell's Deceptive Desires," *Philosophical Quarterly* **36**: 223-229.

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.

Mele, A. R. (1997) "Real self-deception," *Behavioral and Brain Sciences* **20**: 91-136 (with open peer commentary).

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

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

Pinker, S. (1997) *How the Mind Works*, New York, W. W. Norton & Company.

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.

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

Tooby, J., and L. Cosmides (1995) "Foreword," in Baron-Cohen (1995): xi-xviii.

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

Van Leeuwen, D. S. N. (forthcoming) "The Spandrels of Self-Deception," *Philosophical Psychology*.

Response to reviewer on “Finite Rational Self-Deceivers”

The reviewer finds some typos and offers two suggestions for addressing further issues. I correct the typos in the version of the manuscript now being submitted for publication. I do not at length address the two further issues raised for reasons of space. I do, however, note the issues in new endnotes and suggest how they should be addressed. Below I paste the reviewer’s comments and then include the text of each endnote that addresses it.

Reviewer’s comment:
>	(1) On the author's account, self-deception is blind. It is an accidental byproduct of the operation of mental processes that were selected for other effects. But self-deception seems to be more selective than the author acknowledges. For example, people do not usually self-deceive about things that would be imminent threats, even though it would obviously be much more pleasurable not to believe that one faces an imminent threat. The author does distinguish between beliefs that are nearer the sensory periphery and those that are farther from the periphery. But the problem is that some beliefs near the periphery can produce great anxiety, so wouldn't we expect evidence for those beliefs to trigger self-deceptive responses (even if they are ultimately ineffectual)?

Endnote 20 in new manuscript:
20 The fact that evidence needs to be mixed in order for self-deception to happen is crucial for explaining a very important feature of self-deception. Self-deception very rarely occurs concerning perceptual beliefs; in occurs much more often concerning what might be called intangibles. The reason for this is that evidence for or against perceptual beliefs is mixed far less often.

Reviewer’s comment:
>	(2) The author distinguishes three kinds of self-deception, but the author's analysis of the phenomenon of self-deception as a spandrel only seems to apply to the first two, wishful and willful self-deception. It is a puzzle why the mechanisms that the author identifies would generate the third, dreadful self-deception. For example, if Othello really desired that -p (that Desdemona not be unfaithful), we would expect that evidence that p would make him uncomfortable, so how could he be motivated to seek out evidence that p (i.e., evidence that Desdemona is unfaithful). A full account of the capacity for self-deception as a spandrel would have to explain these cases also.

Endnote 28 in new manuscript:
28 Extending this model to cases of willful and dreadful self-deception will involve examining how the kinds of desire constituitively involved in *those* types can trigger the

sorts of selective attention and other processes here identified. For reasons of space, I have not explored such an extension in this paper.

***

I hope you find these changes satisfactory.

Sincerely,
The Author