compos mentis

A Psychological Account of the Formation of Self Deceptive Beliefs

Benjamin Cline
Bethel University

ABSTRACT
The phenomenon of self-deception has puzzled thinkers for centuries. How can this entity known as "the self" come to believe a proposition when it appears to possess sufficient evidence to suspect the proposition's falsity? It is a puzzling phenomenon, and the discussion stemming from self-deception has spawned numerous theories and sub-discussions. Notably, theorists divide over the large questions of intentionality, rationality, and the cohesiveness of the self in the light of self-deception. Traditionally, thinkers have relied heavily on rhetoric and intuition to formulate and defend their theories. In recent years, however, the fields such as cognitive psychology and neurophysiology have begun to contribute critical empirical evidence to the conversation, allowing theorists to put some meat behind their models. Studies in both fields uncover mechanisms hypothesized to play a role in self-deception, providing the groundwork for piecing together a holistic, empirically-advised model of self-deception. This paper works to accomplish just that, integrating cognitive and physiological studies into a coherent model that effectively describes self-deception from an empirically plausible framework. It will first describe a cognitivist model of the phenomenon, positing that the human brain aims first and foremost to create a coherent account of the world around it as quickly as possible. It further hypothesizes that, in the name of survival, the brain accepts the likelihood of minor errors as a consequence of fast information processing in exchange for greater assurance that it will not commit critical errors. This theory uncovers a key point—that the brain is less concerned with how events truly transpire in comparison to the coherence of the brain's narrative and the efficiency of its processing. With the cognitivist model in mind, the paper will explore studies surrounding emotional processing. These physiological findings suggest that the brain's processing of emotionally salient stimuli occurs in parallel with dry, rational cognitive processing. Moreover, it suggests that these parallel streams of processing combine to affect the brain's final output. The paper will conclude by suggesting cognitive dissonance as the underlying initiator of the afore-mentioned information processing biases that lead to self-deception. This claim is made because cognitive dissonance appears to arise out of both a rational and an emotional revulsion to the truth claim being realized. The resulting model aims to provide a holistic, psychologically-informed account of how and why self-deceptive beliefs could arise.

KEYWORDS
Self-deception, cognitive psychology, neurophysiology, emotion, cognitive dissonance, empiricism, psychology, error minimization (PEDMIN), unintentional, information processing, cognitive bias

The phenomenon of self-deception has inspired a great deal of literature. Philosophers and psychologists alike speculate about whether self-deception is intentional or unintentional, whether it is a species of irrationality, and which accounts of "the self" best cohere with the mechanisms of self-deception. Many of these accounts are focused on conceptual analysis, concerned with defining the terms of the debate, generating necessary and sufficient conditions for the meanings of theoretical terms like "intention" or "self." While such conceptual work is foundational, more input from literature steeped in empirical study would advance the discussion. Fortunately, there is a growing body of work that explores the philosophical implications of self-deception from an empirical foundation, using modern psychological and neurological research. This paper will focus on incorporating some of the cognitive and neurological models aimed at explaining self-deception. It will integrate empirical philosophical accounts as well, seeking to create a holistic, empirically-informed model of the mechanisms of self-deception.

Useful Deceptions in Perception

Self-deception, as a primarily neurological phenomenon (that is, a state whose genesis is in the brain), first requires a conceptual framework of the brain and its interaction with the world around it. Perceptual studies of the brain inform us that the brain primarily functions to paint a coherent picture of the external world—an observation evidenced by a variety of optical, auditory, and tactile illusions. These illusions arise out of the brain's tendency to employ heuristics, which are essentially cognitive shortcuts that reduce processing time. This tendency is perhaps most apparent in the brain's visual system, as it employs a small army of heuristics in order to quickly identify pertinent information regarding the surrounding environment. These are exemplified in the Gestalt principles by which the brain organizes and groups objects, using fast perceptual information to discern what part of the scene is the figure and what part is the background. For instance, when presented with an ambiguous picture, the brain often perceives the objects in the bottom of the scene as the figure because, more often than not, this sort of perceptual organization holds true when we interact with the world. This inference allows the brain to quickly make sense of the environment with a remarkably high accuracy. These strategies are not perfect, but they result in quick interpretations of visual information and, ultimately, a quicker physical response to stimuli.

The quickened response made possible by heuristics may determine the difference between the life and death of an organism. When a squirrel decides to scamper up a tree due to a perceived threat, it rarely initiates that action based on complete perceptual information. Rather, it employs heuristics and sacrifices objective accuracy regarding the perceived threat in favor of making a fast decision. Often times, these perceptual shortcuts result in a false alarms, yet these false alarms are regarded as acceptable because one instance of erroneous inaction could end the life of the squirrel. Therefore, perceptual heuristics are vital to the survival organisms. The reinforcing lesson learned from the brain's use of heuristics is that the brain consistently sacrifices an accurate perception of the outside world in favor of forming a coherent picture. This fundamental premise provides a useful foundation to build a theory of self-deception from. This paper will continue to provide evidence in support of this foundational concept, building a model of self-deception around it.

A Cognitive Model of Self-Deception

Cognitive research in psychology provides vast insights into the mechanisms employed by the brain to vet information and make decisions on sensory input. James Friedrich provides an empirical review of these mechanisms, and the resulting analysis has become an oft-referenced cognitive model of psychological mechanisms geared toward self-deception. He builds from the same foundation proposed above, that "our inference processes are first and foremost pragmatic, survival mechanisms and only secondarily truth detection strategies" (Friedrich 1993, 298). What does Friedrich mean by "pragmatic"? To start with, he notes that the brain seeks maximum efficiency by balancing the quality of its information processing with the amount of cognitive effort this processing requires. This proposition does not imply that the brain functions solely to conserve cognitive energy. That would be too simplistic and would undermine both our intuitions regarding the intricacy of the brain's abilities and studies supporting these intuitions. Friedrich hypothesizes that the brain seeks efficiency by working to accomplish a more complex goal than simple energy conservation. He proposes that the brain balances cognitive effort and truth-processing in order to most effectively reduce critical errors. He posits that the greatest danger in decision making is a critical error that results in harm being inflicted upon the organism. Moreover, an effective avoidance of critical errors necessitates the allowance for

smaller errors. For example, in the woods at night, it is more advantageous to perceive a rustling in the underbrush as a predator, even though the overwhelming probability suggests that it is a small animal, the wind, a branch falling, etc… The mantra, "better safe than sorry", applies here, as the brain sacrifices the probability of truth detection (the likely cause of the noise) in favor of avoiding a critical error (falsely assuming there is not a bear in the forest when, in fact, there is). Therefore, the first core proposition of this analysis of self-deception is that humans are pragmatists who are "more concerned with error reduction than truth detection," a proposition evidenced by the perceptual heuristics employed by the brain to come to fast conclusions (Friedrich 1993, 300).

Friedrich continues by offering empirical support of his claim. The resulting picture is a rather intuitive, cohesive method of viewing the brain and its interaction with the surrounding world. For instance, a study found that a certain agency initially judged applicants on the criteria of extraversion/introversion. They were seeking applicants who were friendly, outgoing, and team players. Although there is no true correlation between extraversion and these qualities, the hypothesis exists that extraverts will display these qualities at a higher rate. Therefore, the agency proceeded to cut all applicants who displayed signs of introversion so as to avoid the costly error of hiring a withdrawn, isolated employee (Friedrich 1993). This quick heuristic likely eliminated worthy employees, but it also accomplished the general task of eliminating employees that did not fit the company's mold. Another study asked participants to evaluate hypothetical scenarios where a baked cake turned out either well or poorly, and they were asked to assess the variables that could have caused either the positive or negative outcome. When the cake turned out poorly in the scenario, participants tended to recommend "logically disconfirming (-H)" tests, meaning that they eliminated suspected causes while keeping other, less suspicious variables. When the cake turned out well, though, they "shifted toward +H tests (keeping the suspected cause and eliminating others)" (Friedrich 1993, 302). Friedrich notes that -H tests are equally appropriate for detecting errors. This study, though, showed a change in test strategy when the outcome was positive. Friedrich interprets this as evidence in favor of a PEDMIN analysis, noting that "falsification logic still requires elimination of the suspected causal element, whereas error minimization logic suggests a +H test" (Friedrich 1993, 302).

This account of the brain and its interaction with the world around it suggests a plausible, functional account of self-deception. In this model, the large majority of self deceptive cases could be explained as the output of an organism aimed at attaining maximum efficiency in decision-making processes. To accomplish this, the brain employs methods aimed at reducing costly errors rather than aiming to uncover the truth of how things actually are. These biasing mechanisms could result in altered information encoding, memory suppression, and biased evidence gathering—all harbingers of self-deceptive beliefs. Indeed, the self may be deceived through these processes, but this deception is not a deliberative act. Friedrich's PEDMIN model suggests that it occurs in the realm of fast, subconscious information processing, skewing the data so as to lead to safer, yet biased decisions. Self-deception, then, is portrayed as the outcome of a pragmatic, evolutionarily-advantageous set of processes that protects the organism, sacrificing accuracy of incoming information in favor of avoiding situations where critical errors may occur.

Emotional Processing in Self-Deception

From the evidence presented above, the proposition that people process information with the primary intent of avoiding costly errors appears to present a valid picture of how the brain works. Certainly, cognitive biases can skew perceptions of reality, and a pragmatic strategy for information processing and error identification gives reason to skew reality. However, the PEDMIN analysis misses a key aspect of information processing—the emotional aspect. Friedrich notes that the PEDMIN model does not imply that the brain consistently carries out its error-reducing functions accurately and correctly, but he does not make any move to include other major factors that could skew the brain's analysis of information. There are diverging accounts from Friedrich's suggesting that humans do not process data from a purely logical standpoint. They point to cases of self-deception that do not appear to be grounded solely in skewed, pragmatic processing mechanisms, suggesting that emotionally-biased information processing influences the state. This is corroborated by psychological and physiological evidence showing that humans are not fully efficient or pragmatic on either the conscious level or the unconscious. Though pragmatism may account for a piece of a theory of self-deception, emotional processing appears to also play a role in the full process.

There are a number of accounts that incorporate emotion into theories of self-deception. They explore how the intensity of emotion can destabilize a person's rationality and motivate self-deceptive beliefs. These accounts approach from a variety of directions—conceptual, computational, neurological, and psychological—helping to flesh out the concept and to aid in filling the apparent void where the PEDMIN analysis falls short (Correia, 2014; Sahdra and Thagard, 2003; Halgren and Marinkovic, 1995; Scott-Kakures, 2009.). It should first be noted that an integration of emotional processing into a PEDMIN analysis of information processing does not contradict the PEDMIN theory. It does not require that we weaken the claims of the theory, either. Instead, a deeper understanding of the role played by emotional processing in facilitating self-deception integrates itself into the existing conception put forward, creating a more complete picture of the phenomenon. The available evidence will be evaluated in the following section and integrated into this emerging picture.

In the computational realm, two researchers created two differing models of information processing aimed at explaining the genesis of self deception (Sahdra and Thagard, 2003). They specifically used their models to explore how Dimmesdale, the adulterous minister in "The Scarlet Letter," could have processed the conflicting information of his sins interposed on his role as a spiritual leader. They present the "Cold Clergyman" and the "Hot Clergyman," delineating between a cold, rational self-deceptive analysis and a hot, emotionally-steeped descent into self-deceptive beliefs. In the rational analysis, the primary goal sought was coherence. When two incoherent propositions appear in this model, other third-party propositions check the first two, weighing all factors to create the most rational, coherent set of propositions. This sort of analysis runs similarly to the PEDMIN analysis. In their second, "hot" analysis, the researchers introduced emotional valences into their computational neural networks. The positive or negative values of the valences influenced the rational propositions, changing their weights. In the final version of the researchers' model, emotional valences altered the way that information was processed and led to different self-deceptive propositional outcomes when run on the propositional web of Dimmesdale's self-deception. By showing that the inclusion of emotions in the processing of information alters the propositions involved in a self-deceptive belief system, the authors show the possibility that emotional processing is involved in the formation of beliefs.

The computational conclusions put forward are corroborated by extraordinary findings in the physiological realm. This paper proposes that emotional processing occurs in the same fast, unconscious processing realm as the information processing put forward in the PEDMIN analysis. An EEG study by Eric Halgren and Ksenija Markinkovic supports this proposition. In their study, they recorded electrical signals from participants during the processing of emotionally charged stimuli. Their EEG readings show limbic system activation beginning 120 ms after stimulus onset. The limbic system contains structures associated with emotional processing, suggesting that this processing begins early on in information processing. Moreover, this sort of timeline suggests that it occurs concurrently with other fast cognitive processing. Such a synchrony of processes "permits limbic input to shape the content of the encoded experience rather than simply to react to its content" (Halgren and Markinkovic 1995, 1146). The importance of this point should be underscored, as it provides evidence that emotional processing is an *agent* in determining the encoded experience rather than an *outcome* of processing from different mechanisms. Since emotional processing occurs so early in the processing of information, it likely influences the outcome rather than simply reacting to the outputs of other processes. Moreover, this means that emotional processing stemming from the limbic system of the brain could contribute to "the myriad of psychological defense mechanisms that may distort or eliminate the conscious experience of an emotionally significant event" (Halgren and Markinkovic 1995, 1146). This direct physiological evidence of fast, emotionally salient processing provides tantalizing evidence in support of the role of emotion within the brain's analysis of information. The model beginning to arise from this evidence suggests that the information-processing biases instigated by emotionally salient stimuli function below the realm of conscious experience. This sort of parallel processing allows for emotions to factor into the final analysis of information while still allowing for the pragmatic PEDMIN analysis to occur separately.

Integrating a Cognitivist Approach with Emotional Processing

A slew of authors have worked to incorporate aspects of emotional processing into a complete picture of self-deception, and they fall at different points across the spectrum regarding the pervasive questions of self-deception. For instance, thinkers debate the intentionality underlying self-deceptive states when emotion

comes into play. Some thinkers understand the phenomenon as an entirely unintentional process (Mele, 2003; Correia, 2014). The prevailing consensus across such accounts posits that self-deception arises out of some sort of cognitive bias (or biases), where the self-deceived is a "victim of a phenomenon of judgement distortion that is both involuntary and unconscious" (Correia 2014, 317). While an unintentional view of self-deception tends to prevail when incorporating emotions, certain accounts hedge on the answer and do not provide a solid move toward intentionality or unintentionality. Nelkin proposes that the "desire to believe" is a necessary condition of self-deception. This desire, however, "need not be conscious," leaving the question of intentionality open to situational influence (Nelkin 2002, 395). In this instance, the author argues that—though the individual is likely unaware of the actual biasing process or of the biasing effect that emotional processes have on information—the individual must have an intentional, motivational hand in initiating the unconscious process.

Nelkin's account is possible. However, the current model rejects an intentional account based on the psychological and physiological evidence presented above. While an individual may experience the feeling of direct control over his emotions, research such as Halgren's suggests that the genesis of such thoughts is in the fast emotional processing occurring in the limbic system, which resides outside the realm of conscious control. The clash between the two hypotheses leads to a circular debate regarding the genesis of the process. Does the desire direct neural processing or does neural processing direct the desire? When stated in more psychological terms, the question of intention turns into a debate between top-down and bottom-up processing as the instigator of the process. For this reason, we will eschew the terms "intentional" and "unintentional" for now in favor of the more psychologically relevant terms. This top-down/bottom-up debate is still contested. For the purpose of this paper, it could perhaps be sidestepped succinctly by ensuring that we hold to a tight definition of self-deception. If one is not careful, the concept of self-deception can slip into the realm of simple wishful thinking, the genesis of which is more obviously a top-down phenomenon. Suffice it to say, the interpretation of the psychological and physiological evidence presented above converges to provide a reasonable explanation of the phenomenon as the output of a fully unconscious, bottom-up process.

Mele defends a similar hypothesis to the one stated above. He posits the following thesis:

> In some instances of entering self-deception in acquiring a belief, an emotion makes a biasing contribution to the production of that belief that is neither made by a desire nor causally mediated by a desire. (Mele 2003, 168).

He also holds the opinion that emotional processing can bias the acquisition of a belief, which mirrors the earlier suggestions from the earlier studies. Furthermore, Mele integrates a PEDMIN analysis into the emotional process. He adds a wrinkle to PEDMIN, though. Given one's emotions about a particular state (e.g. I am fearful my wife is cheating on me), the *emotional state* resulting from evidence confirming (or denying) the proposition is in itself a costly error. Therefore, this iteration of the PEDMIN model incorporates emotional states in that emotional states are factors and consequences in the error analysis, serving as both costly and non-costly outcomes to weigh. This account is attractive in the way that it integrates PEDMIN and emotional processing. Trivers adds to this assertion. He brings the reader's attention to a plethora of psychological studies showing that humans tend to encode information in a positive light, at times entirely failing to encode material that evokes negative emotions about the self (Trivers 2011). He shows that the brain's encoding of events is biased by positive or negative affect to the point that it causes self-deceptive recollections of that event. Even seemingly dry, emotionless instances of self-deceptive beliefs do not escape some level of implicit emotional biasing. For instance, simple PEDMIN perceptual errors (which could hardly be counted as self-deceptive) still activate an emotional response regarding the potential cost of significant errors. For example, the possibility of allowing physical harm to be done upon myself by ignoring the rustling I hear in the forest at night will bring about emotions like fear, which will influence my processing of the visual and auditory information I am perceiving. It appears that emotions are attached in even the most modest of perceptual biases.

Cognitive Dissonance: A Motivating Initiator for Self-Deceptive Processing

From the previous arguments, we are given a picture of self-deception as an integration of the brain's rational function to minimize costly errors alongside the biased processing of emotionally salient information. It appears that these processes initially function independently, but their combined effect results both in biased decision making and inaccurate encoding of information. Within this understanding, a key element is missing that has been mostly ignored to this

point—motivation. This account maintains that the motivation for self-deception is entirely unconscious and unintentional. In previous sections, it has only briefly discussed the motivation underlying the processes—survival in the case of PEDMIN and a sort of emotional coherence or avoidance of noxious stimuli in emotional processing. These vague understandings should be explored more. In accounting for these underlying motivation, a remarkable parallel with the social psychological phenomenon of cognitive dissonance emerges. The final section of this paper will integrate cognitive dissonance into the previously established understanding of self-deception, exploring its role as an unconscious motivator for both the PEDMIN and the emotional processes.

Cognitive dissonance can be used both to explain the motivation underlying self-deceptive processes and to give an account of how people are able to vehemently defend seemingly dubious beliefs. The theory of cognitive dissonance posits that, first and foremost, human beings strive for consistency (Festinger 1957). When discrepant cognitions and actions appear, they produce an uncomfortable psychological state. For instance, my preference to view myself as as an honest individual conflicts with the white lie I told to my uncle to avoid an awkward confrontation in the family. This uncomfortable state results in the selection and pursuit of dissonance-reducing strategies. These dissonance-reducing strategies—behaviors such as thought suppression, biased evidence seeking, and biased information encoding—lead to self-deceptive beliefs. Scott-Kakures adds to the traditional account of cognitive dissonance, noting that humans must spend a large amount of energy on settling questions about reality when they process the events in detail, making it advantageous to eliminate discrepant cognitions before they are fully processed (Scott-Kakures 2009). Moreover, it is less cognitively taxing to come to gain and maintain certainty about conclusions, even within a clearly uncertain environment. As the PEDMIN model noted above, the brain cannot waste time ascertaining the true danger of a costly error within a situation. Rather, it best serves by asserting a confident perception of reality that most effectively minimizes the risk of critical errors.

With the previous points in mind, a summary of the proposed model of this paper is as follows. It posits cognitive dissonance as the motivation that underlies the parallel process that the brain uses to mediate and eliminate this dissonance. These processes initiate a PEDMIN analysis that is influenced and adapted by simultaneous emotional processing. The brain employs this strategy in order to

eliminate cognitive dissonance, allowing for minor misinterpretations of data in order to avoid the greater consequences of making larger errors. These errors can be both practical and emotional, as it is disadvantageous both to be eaten by a bear and to lose positive emotions regarding one's self-conception. These misinterpretations of data are further driven by the emotional valences both of the information being processed and the propositions about the self that are at stake as a result of the analysis. This self-deceptive analysis happens quickly, and it occurs with speed for a handful of reasons. First, as Trivers and Scott-Kakures mentioned, the sooner the brain eliminates discrepant information—either through biased encoding, directed forgetting, or thought suppression—the less cognitive resources the process of elimination consumes (Trivers 2011; Scott-Kakures 2009). Oftentimes, this analysis works so quickly and effectively that cognitive dissonance never arises, decreasing the amount of effort required to suppress the inherent contradictions in self-deceptive beliefs. Regardless of the timeline, though, the model employs these processes to alleviate the negative effects of cognitive dissonance.

This model of the processes that facilitate self-deception stands on an empirical foundation. Individually, PEDMIN-like cognitive processing, emotional processing, and cognitive dissonance have all garnered the widespread support of empirical literature. The combination of these processes creates an plausible, integrative model of how self-deceptive beliefs may arise in humans. Due to the literature supporting its individual components, this work is more than simply conceptual and speculative. It represents a psychologically plausible model of the phenomenon and should stand the test of conceptual criticism due to its supporting literature.

References

Correia, Vasco. 2014. "From Self-Deception to Self-Control: Emotional Biases and the Virtues of Precommitment." *Croatian Journal of Philosophy* 14 (42): 309-323.

Festinger, Leon, 1919-1989. 1957. *A Theory of Cognitive Dissonance*. Evanston, Ill: Row, Peterson.

Friedrich, James. 1993. "Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena." *Psychological Review* 100 (2): 298-319. doi:10.1037/0033-295X.100.2.298.

Halgren, Eric, and Ksenija Marinkovic. 1995. "Neurophysiological Networks Integrating Human Emotions." In *The Cognitive Neurosciences*, edited by Michael Gazzaniga, 1137–1151. Cambridge, MA: MIT Press.

Mele, Alfred R. 2003. "Emotion and Desire in Self-Deception." *Royal Institute of Philosophy Supplement* 52: 163-179.

Nelkin, Dana K. 2002. "Self-Deception, Motivation, and the Desire to Believe." *Pacific Philosophical Quarterly* 83 (4): 384-406.

Sahdra, Baljinder, and Paul Thagard. 2003. "Self-Deception and Emotional Coherence." *Minds and Machines* 13 (2): 213-231.

Scott-Kakures, Dion. 2009. "Unsettling Questions: Cognitive Dissonance in Self-Deception." *Social Theory and Practice* 35 (1): 73-106.

Trivers, Robert. 2011. *The Folly of Fools.* New York: Basic Books.