Self-Deception and Emotional Coherence

BALJINDER SAHDRA and PAUL THAGARD
University of Waterloo, Department of Psychology, Faculty of Arts, 200 University Avenue West, N2L 3G1, Waterloo, Ontario, Canada

Abstract. This paper proposes that self-deception results from the emotional coherence of beliefs with subjective goals. We apply the HOTCO computational model of emotional coherence to simulate a rich case of self-deception from Hawthorne's *The Scarlet Letter.* We argue that this model is more psychologically realistic than other available accounts of self-deception, and discuss related issues such as wishful thinking, intention, and the division of the self.

Key words: coherence, desire, emotion, self, self-deception, simulation, wishful thinking

1. Introduction

Skeptics such as Paluch (1967) and Haight (1980) think that the very notion of self-deception is implausible. However, there is empirical evidence that self-deception is not only possible but also highly pervasive in human life. It accounts for positive illusions of opponents in battle and their belief that they will win (Wrangham, 1999). It is involved in denial in physical illness (Goldbeck, 1997). It has been shown to account for unrealistic optimism of the self-employed (Arabsheibani et al., 2000). It has been observed in traffic behavior of professional drivers (Lajunen et al., 1996). And it has been shown to mediate cooperation and defection in a variety of social contexts (Surbey and McNally, 1997).

What is self-deception? How do we deceive ourselves? Researchers have attempted to answer these questions in various ways. Some thinkers argue that self-deception involves a division in the self where one part of the self deceives the other (Davidson, 1985; Pears, 1986; Rorty, 1988, 1996). Others, however, maintain that such division is not necessary (Demos, 1960; Elster, 1983; Johnston, 1988; McLaughlin, 1988; Rey, 1988; Talbott, 1995; Mele, 2001). Some consider self-deception to be intentional (Sackeim and Gur, 1978; Davidson, 1985; Pears, 1986; Rorty, 1988; Talbott, 1995), while others insist that it is non-intentional (Elster, 1983; Johnston, 1988; McLaughlin, 1988, 1996; Lazar, 1999; Mele, 2001). Some think that self-deception is a violation of general maxims of rationality (Pears, 1982; Davidson, 1985), while others argue that self-deception is consistent with practical rationality (Rey, 1988; Rorty, 1988).

We propose that self-deception can result from emotional coherence directed to approach or avoid subjective goals. We will show this by modeling a specific case of self-deception, namely that of Dimmesdale, the minister in Hawthorne's *The Scarlet Letter* (1850). The model has two parts, namely, a "cold" or non-emotional

Minds and Machines **13:** 213–231, 2003.

model and a "hot" or emotional model. The emotional aspect of self-deception may be implicit in some accounts, but no one has brought it in the forefront of the discussion. Two notable exceptions are Ronald de Sousa (1988) and Ariela Lazar (1999). Our computational account is more precise than de Sousa's. Lazar (1999) argues that self-deceptive beliefs are partially caused by emotions whose effects are not mediated by practical reasoning. Our account differs from Lazar's in that we show that self-deception can arise even in those cases in which self-deceivers are highly motivated to be rational; the effects of emotions are just as mediated by rational thought as the effects of rational thought are by emotions. In other words, we will show that it is the interaction of cognitive and emotional factors that plays the pivotal role in self-deception.

After giving our model, we will also compare it to two other models of self-deception, namely Rey's (1988), and Talbott's (1995). We will argue that our model is more psychologically plausible than their models.

2. What Is Self-Deception?

Self-deception involves a blind or unexamined acceptance of a belief that can easily be seen as "spurious" if the person were to inspect the belief impartially or from the perspective of the generalized other (Mitchell, 2000, p. 145). Consider an example from Mele (2001, p. 26): Don deceives himself into believing the belief p that his research paper was wrongly rejected for publication. Indubitably, there are two essential characteristics of Don's self-deception:

1 Don *falsely* believes p.
2 Either someone other than Don, or he himself at a later *impartial* examination, observes (or accuses) that he is deceiving himself into believing p.

The two points are related. The reason we know that Don falsely believes p is that an impartial examination of evidence suggests that Don ought to believe $\sim$p. Most of the time, "the impartial observer", to use Mele's (2001) terms, is someone other than the self-deceiver, but it is also possible that the self-deceiver herself may realize the spuriousness of and self-deception involved in her beliefs, at a later careful examination.

In 2 above, one might even use the term "interpretation" instead of observation or accusation. Goldberg notes, "accusations of self-deception are only as strong as the interpretation of . . . [the] behavior" (Goldberg, 1997). Thus, one might say that the model of Dimmesdale's self-deception that we will present shortly is based on *our* interpretation of his speech and beliefs as they are revealed in the narrative. But we cannot have just any interpretation. The best interpretation is the one that is consistent with all the available information.

The minimal description of self-deception that we have given above is not finely tuned to distinguish self-deception from wishful thinking and denial. We will make

such distinctions after we give a much more precise account of self-deception in our computational model. Given these remarks, we can begin the impartial or external analysis of the self-deception of Dimmesdale, the minister in *The Scarlet Letter*.

3. Dimmesdale's Self-Deception

Before we present our analysis, we must clarify that our purpose is not to morally judge Dimmesdale. Some theorists hold that self-deception is intrinsically wrong in that it is a sort of spiritual failure (Sartre, 1958; Fingarette, 1969). At the same time, many philosophers also argue that self-deception is not always wrong, and may even be beneficial; for example, see Rorty (1996) and Taylor (1989). Nevertheless, in the era in which Hawthorne wrote *The Scarlet Letter*, it was taken for granted that being dishonest with oneself was somehow always wrong. Hawthorne gives us the famous advice: "Be true! Be true! Be true! Show freely to the world, if not your worst, yet some trait whereby the worst may be inferred!" (Hawthorne, 1850, p. 260) For Hawthorne, self-deception, like hypocrisy, is morally wrong because self-deceivers are largely out of touch with their true selves within them and mistakenly place their trust in what the reader recognizes to be a false outer appearance (Harris, 1988, p. 21). We think that the issue of moral reprehensibility of self-deception is important. However, our objective is much humbler in that we only aim to *explain* how Dimmesdale deceived himself, and thus suggest a new way of conceiving of self-deception.

The Scarlet Letter is particularly attractive for a study of self-deception because of the richness of detail of the self-deceiving characters in it. This degree of detail is invariably missing in the philosophical literature on self-deception; the most commonly cited case is Elster's (1983) example of sour grapes. In *The Scarlet Letter*, almost everybody is a hypocrite. More importantly, Hawthorne's hypocrites are almost always self-deceivers as well. The reader easily infers that from the narrative. On one occasion, however, the narrator explicitly informs us that Dimmesdale, the minister, had deceived himself; and later, Dimmesdale himself unfolds his character to reveal the complexity of his self-deception. See Figure 1 for a pictorial analysis of Dimmesdale's self-deception. The solid lines in the figure represent positive associations, and the dotted lines represent negative associations.

The narrator informs us that Dimmesdale deceives himself when he tells himself that his satisfaction at knowing he can still preach the Election Day sermon before running off with Hester arises from his wish that people will think of him as an admirable servant of the Church. He says to Hester: People "shall say of me . . . that I leave no public duty unperformed, nor ill performed" (Hawthorne, 1850, p. 215). "Sad, indeed," the narrator tells us, "that an introspection so profound and acute as this poor minister's should be so miserably deceived!" (p. 215) Dimmesdale deceives himself in not wanting to have it said that he failed to carry out his "public duty," even if it were to be revealed almost immediately, as it obviously would, to

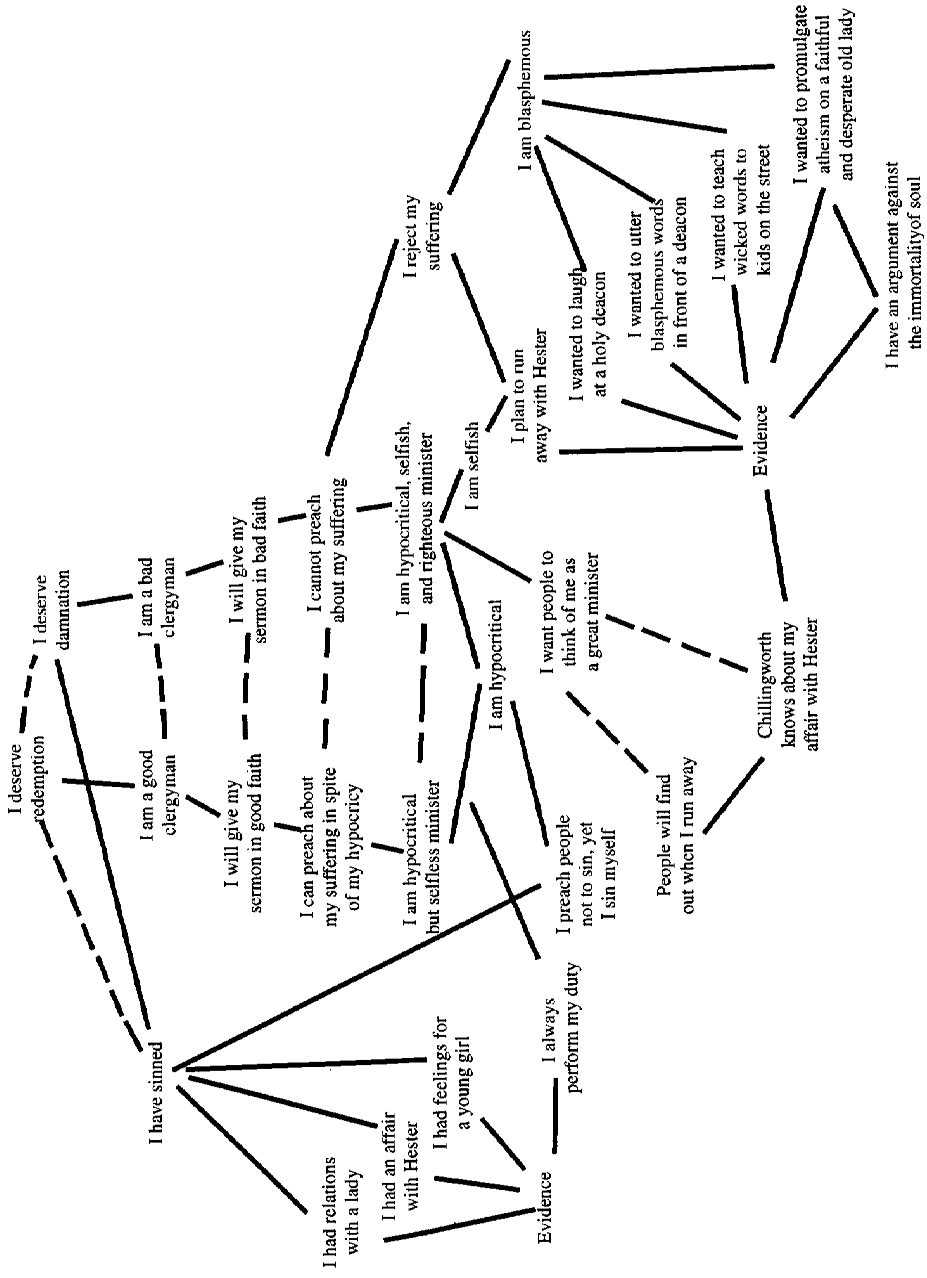

Figure 1. ECHO analysis of Dimmesdale's self-deception. Solid lines represent positive associations. Dotted lines represent negative representations.

the same public that his dutiful service to them was sheer hypocrisy. In other words, his self-deception consists in believing that he can fill his sacramental office and still be hypocritical.

Dimmesdale deceives himself in believing that he is no more hypocritical now than he has been for seven years. Previously, he has known that he has been hypocritical because of his hidden sinfulness. However, at one point during his conversation with Chillingworth, he concedes that a priest who knows he is guilty may still have the obligation to continue as a priest, and in that sense he would not necessarily be a hypocrite. He claims that some people, "by the very constitution of their nature," may be forced to live with the "unutterable torment" of hidden guilt and still carry out ministerial, even sacramental functions (p. 132). Therefore, as Harris puts it, "in the past, Dimmesdale has been a good clergyman not only in spite of his hidden guilt and his consciousness of his hypocrisy, but precisely *because* of those very factors — because of his excruciating situation" (Harris, 1988, p. 84). He has been a hypocrite because he has allowed people to think that he is a saint; but his motive in doing that has been to fulfill his duty. Thus, he was a good clergyman in spite of his hypocrisy because his motives were selfless.

The situation is different however, when he deceives himself. His motives have changed and he deceives himself in believing that his motives are the same. This time, his motive is to pass himself off as righteous. His main concern is that people say of him that he "leaves no public duty unperformed, nor ill performed!" (Hawthorne, 1850, p. 215). In the past, Dimmesdale has been a hypocrite but still a good clergyman. Now, he is a hypocrite and a bad clergyman because his motives are selfish. He believes that he can still sincerely preach about his suffering even though he rejects his suffering now. In short, he deceives himself in believing that he is a good clergyman now as he was in the past.

The fact that Dimmesdale's self-deception is intertwined with his hypocrisy causes him much confusion. Harris thinks that his self-deception is of the "deepest, most unconscious sort, compounded by deliberate hypocrisy, and the prognosis calls for an ever-increasing confusion of identity" (Harris, 1988, p. 75). The narrator in the novel describes this complexity: "No man, for any considerable period, can wear one face to himself, and another to the multitude, without getting bewildered as to which may be the true" (Hawthorne, 1850, pp. 215–216). In the chapter, "The Minister in a Maze," as Dimmesdale's identity unravels, his impulses seem to be "at once involuntary and intentional: in spite of himself, yet growing out of a profounder self than that which opposed the impulse" (p. 217). This "profounder self" incites him into blaspheming in front of a deacon; laughing at the good and holy deacon; promulgating atheism to a helpless old lady who has nothing but faith in God to sustain her; teaching "some very wicked words" to a few kids playing on a street; and giving a "wicked look" to a young girl with a spiritual crush on him.

His "profounder self" has much to do with his sexuality as manifested in three things: First, his impulsive affair with Hester; second, his likely relations with "the

many blooming damsels, spiritually devoted to him" from his congregation (p. 125). Third, as mentioned previously, when he meets one young girl while he is lost in his private "maze," the reader is informed, "that the minister felt potent to blight all the field of innocence with but one wicked look" (p. 220).

It is important to note that Dimmesdale is able to end his self-deception. This will factor significantly in our model, as we will explain later. Throughout the novel, Hawthorne is very sarcastic and scornful of Dimmesdale for his hypocrisy and self-deception, but in the end he makes him appear as a kind of a hero or saint. Dimmesdale, while giving the sermon, "stops deceiving himself into thinking that he could preach about his suffering at the same time he was planning to reject his suffering" (Harris, 1988, p. 86). He preaches in the same spirit as before and to the same effect. Thus, he returns to his earlier state of hypocrisy. He escapes his hypocrisy at the time of his death when he declares it in front of all the people of his congregation. The reason he is able to escape his self-deception and his hypocrisy is that he knows, more than anybody else, that he is unworthy of redemption.

4. General Description of Our Model of Dimmesdale's Self-Deception

We used the simulators, ECHO and HOTCO 2 to computationally model Dimmesdale's self-deception. The two sections following this one contain the detailed descriptions of these simulators. In this section, we give a general description of our model.

The model has two parts: (1) Cold Clergyman, a cold or emotionless explanation, and (2) Hot Clergyman, an emotional explanation. The first part is the test to see what Dimmesdale *would* believe given the evidence. This experiment would serve as a rational baseline. In other words, this would be the impartial or external observation of the situation. The second part is the model of what he *does* believe in spite of the evidence, given his goals and emotional preferences.

In the first experiment, Cold Clergyman, the input is simply the observed propositions (that is, the evidence), and the negative and positive associations between propositions (see Figure 1). After the experiment is run, we expect that Dimmesdale would believe the belief-set A:

1 I am a bad clergyman.
2 I will give my sermon in bad faith.
3 I cannot preach about my suffering.
4 I am hypocritical, selfish, and righteous minister.

In the second experiment, Hot Clergyman, in addition to the evidence, propositions, and all the associations, he is given two goals: (1) approach redemption, and (2) avoid damnation. Also, he is given 'likes' and 'dislikes', based on whether a proposition has negative or positive emotional valence (See Table 1). For ex-

Table I. Dimmesdale's likes and dislikes

	Propositions	Goals	Likes	Dislikes
1	I am a good minister.		*	
2	I am a bad minister.			*
3	I deserve redemption.	Approach redemption		
4	I deserve damnation.	Avoid damnation		
5	I will give my sermon in good faith.		*	
6	I will give my sermon in bad faith.			*
7	I can preach about my suffering in spite of my hypocrisy.		*	
8	I am hypocritical but selfless.			*
9	I always perform my duty.		*	
10	I cannot preach about my suffering.			
11	I am hypocritical, selfish, and righteous minister.			
12	I reject my suffering.			*
13	I am blasphemous.			*
14	I am hypocritical.			*
15	I am selfish.			*
16	I have sinned.			*
17	I preach people not to sin, yet I sin myself.			*
18	I had an affair with Hester. (E)			*
19	I had feelings for a young girl. (E)			*
20	I had relations with a lady. (E)			*
21	I want people to think of me as a great minister. (E)			
22	People will discover my guilt.			*
23	Chillingworth knows about my affair with Hester. (E)			*
24	I plan to run away with Hester. (E)			*
25	I wanted to laugh at a holy deacon. (E)			
26	I wanted to utter blasphemous words in front of a deacon. (E)			
27	I wanted to teach wicked words to kids. (E)			
28	I wanted to promulgate atheism on an old lady. (E)			
29	I have an argument against the immortality of soul. (E)			

(E) = Evidence.

ample, he likes being a good clergyman and dislikes being a bad clergyman. In this experiment, he should be able to deceive himself into believing the belief-set B:

1 I am a good clergyman.

2 I will give my sermon in good faith.
3 I can preach about my suffering in spite of my hypocrisy.
4 I am hypocritical but selfless.

Cold clergyman is run in the explanatory coherence program, ECHO. Hot Clergyman is run in the emotional coherence program, HOTCO 2. We devote the following two sections to describe ECHO and HOTCO 2 in detail.

5. ECHO and Explanatory Coherence

ECHO is an implementation of the theory of explanatory coherence that can be summarized in the following principles, discussed at length elsewhere (Thagard, 1992, 2000).

- *Principle E1. Symmetry.* Explanatory coherence is a symmetric relation, unlike, say, conditional probability. That is, two propositions p and q cohere with each other equally.
- *Principle E2. Explanation.* (a) A hypothesis coheres with what it explains, which can either be evidence or another hypothesis; (b) hypotheses that together explain some other proposition cohere with each other; and (c) the more hypotheses it takes to explain something, the lower the degree of coherence.
- *Principle E3. Analogy.* Similar hypotheses that explain similar pieces of evidence cohere.
- *Principle E4. Data priority.* Propositions that describe the results of observations have a degree of acceptability on their own.
- *Principle E5. Contradiction.* Contradictory propositions are incoherent with each other.
- *Principle E6. Competition.* If P and Q both explain a proposition, and if P and Q are not explanatorily connected, then P and Q are incoherent with each other. (P and Q are explanatorily connected if one explains the other or if together they explain something.)
- *Principle E7. Acceptance.* The acceptability of a proposition in a system of propositions depends on its coherence with them.

ECHO shows precisely how coherence can be calculated. Hypotheses and evidence are represented by units, which are highly simplified artificial neurons that can have excitatory and inhibitory links with each other. When two propositions cohere, as when a hypothesis explains a piece of evidence, then there is an excitatory link between the two units that represent them. When two propositions are incoherent with each other, either because they are contradictory or because they compete to explain some of the evidence, then there is an inhibitory link between them. Standard algorithms are available for spreading activation among the units until they reach a stable state in which some units have positive activation, representing the

acceptance of the propositions they represent, and other units have negative activation, representing the rejection of the propositions they represent. Thus algorithms for artificial neural networks can be used to maximize explanatory coherence, as can other kinds of algorithms (Thagard and Verbeurgt, 1998; Thagard, 2000).

6. HOTCO and Emotional Coherence

When people make judgments, they not only come to conclusions about what to believe, but they also make emotional assessments. For example, the decision to trust people is partly based on purely cognitive inferences about their plans and personalities, but also involves adopting emotional attitudes toward them (Thagard, 2000, Ch. 6). The theory of emotional coherence serves to explain how people's inferences about what to believe are integrated with the production of feelings about people, things, and situations. On this theory, mental representations such as propositions and concepts have, in addition to the cognitive status of being accepted or rejected, an emotional status called a *valence*, which can be positive or negative depending on one's emotional attitude toward the representation. For example, just as one can accept or reject the proposition that Dimmesdale committed adultery with Hester, one can attach a positive or negative valence to it depending on whether one thinks this is good or bad.

The computational model HOTCO implements the theory of emotional coherence by expanding ECHO to allow the units that stand for propositions to have valences as well as activations. Valences are affective tags attached to the elements in coherence systems. Valences can be positive or negative. In addition, units can have input valences to represent their intrinsic valences. In the original version of HOTCO (Thagard, 2000), the valence of a unit was calculated on the basis of the activations and valences of all the units connected to it. Hence valences could be affected by activations and emotions, but not vice versa: HOTCO enabled cognitive inferences such as ones based on explanatory coherence to influence emotional judgments, but did not allow emotional judgments to bias cognitive inferences. HOTCO and the overly rational theory of emotional coherence that it embodied could explain a fairly wide range of cognitive-emotional judgments involving trust and other psychological phenomena, but could not adequately explain Dimmesdale's self-deception.

Thagard (forthcoming) altered HOTCO to allow a kind of biasing of activations by valences. This version of the program, HOTCO 2, allows biasing for all units. For instance, consider the proposition, "I will be redeemed." This proposition can be viewed as having an activation that represents its degree of acceptance or rejection, but it can also be viewed as having a valence that corresponds to Dimmesdale's emotional attitude toward redemption. Since he deems redemption as of great importance, this proposition has a positive valence. In HOTCO 2, therefore,

truth and desirability of a proposition become interdependent. Technical details concerning explanatory and emotional coherence are provided in an appendix.

7. Results of Cold Clergyman and Hot Clergyman

As expected, in the cold-experiments run in ECHO, the system yields acceptance of all four propositions of the belief-set A:

1 I am a bad clergyman.
2 I will give my sermon in bad faith.
3 I cannot preach about my suffering.
4 I am hypocritical, selfish, and righteous minister.

Also, the system rejects the belief-set B:

1 I am a good clergyman.
2 I will give my sermon in good faith.
3 I can preach about my suffering in spite of my hypocrisy.
4 I am hypocritical but selfless.

On the other hand, in the hot experiments run in HOTCO 2, given that the weight of the input valence is equal to or greater than 0.06, the system successfully self-deceives; that is, the belief-set B is accepted and A is rejected (except for proposition 4 in A). The ideal solution to model Dimmesdale's self-deception, however, is when the weight is 0.07. At this degree of input valence, Dimmesdale successfully deceives himself into believing the belief-set B while he rejects the proposition that he will be redeemed. In other words, self-deception occurs, but the proposition that he will be redeemed has a negative activation, that is, it is rejected. Also, Dimmesdale fully accepts that he has sinned. This is consistent with the novel, *The Scarlet Letter*, in which Dimmesdale never denies that he has sinned and experiences much pain due to his guilt. The result is also consistent with the fact that Dimmesdale is able to escape his self-deception, and admit his sin in front of his congregation before he dies with a cry for forgiveness. Thus, HOTCO 2 successfully models that although Dimmesdale is trying to approach redemption while deceiving himself, he never fully approaches it. This allows him to get out of his self-deception later in the novel when he realizes that he can never be redeemed unless he escapes his self-deception and hypocrisy.

8. Self-deception and Subjective Well-Being

According to our model, self-deception occurs via emotional biasing of people's beliefs, while people attempt to avoid or approach their subjective goals. This account is consistent with Erez et al.'s (1995) psychological theory of subjective

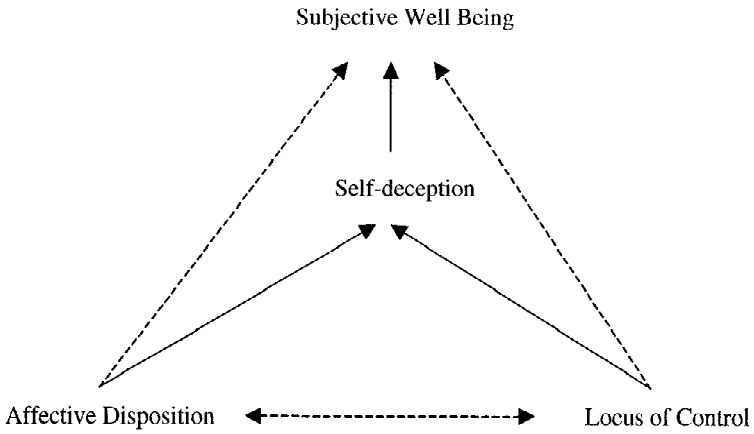

Figure 2. The Psychological Causal Model of Self-deception. Adapted from Erez et al. (1995)

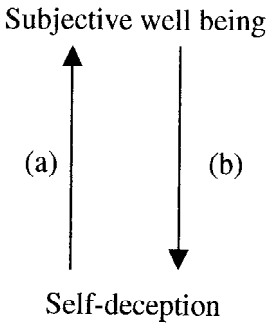

Figure 3. The two-way causal-link between self-deception and subjective well being.

well being according to which dispositional tendencies, such as, affective disposition and locus of control, influence subjective well being through self-deception. According to this theory, certain individuals tend to use self-deception in order to maintain their happiness. Such individuals are either positively disposed or they have high expectations of control. They tend to ignore failure, for instance, if they are positively disposed. They unrealistically think that they control their environment, if they have high expectations of control. Also, individuals who tend to evaluate stimuli in a positive manner or tend to think they can control their environment do so by actively searching for positive and desirable cues while denying negative and undesirable one (Erez et al., 1995) (see Figure 2).

Thus, focusing on the bottom section of the Erez et al.'s hypothesized causal model, two things may cause self-deception: affective disposition and focus of control. However, we hypothesize that there is a two-way causal-link between self-deception and subjective well being (see Figure 3).

We think that the causal-link is bi-directional because:

(a). There is evidence that self-deception causes subjective well being:

- Self-deception is one of the mental mechanism that increases the subjects' positive assessments of situations (ignoring minor criticisms, discounting failure, and expecting success) (Zerbe and Paulhus, 1987).
- Self-deceivers continually distort daily events to build positive esteem (Paulhus and Reid, 1991).
- Self-deception improves motivation and performance of competitors by reducing their stress and bolstering their confidence by repressing their knowledge of the self-interests and confidence of their competitor (Starek and Keating, 1991).

(b). Also, subjective well being causes self-deception:

- The positive esteem, when sufficiently strong, may act as a buffer to soften the impact of negative information (Erez et al., 1995). Thus, having high subjective well being may cause one to self-deceive by causing one to ignore negative information.
- When threatening or "harmful-to-the-self" information is presented to ego enhancers, they turn to their assets and emphasize them to neutralize the threat (Taylor, 1989). Thus, an enhanced ego can cause one's self-deception in that it neutralizes the influence of any evidence that diminishes ego.

In our model described in the previous sections, self-deception is directed toward approaching or avoiding certain subjective goals, which presumably increase or decrease subjective well being if approached or avoided. For instance, in Dimmesdale's case, the causal model can be depicted as shown in Figure 4, which shows that Dimmesdale's subjective well being depends on his prospect of being redeemed. Being a good clergyman is essential for redemption. He may be disposed to ignore any evidence that may suggest that he is a bad clergyman. This is consistent with HOTCO 2 experiments in which the proposition that people will know when he runs away with Hester is rejected. Also, he may have a false sense of control that he will give his sermon in good faith, and make people believe that he is a good minister. His false sense of control and his disposition to ignore certain evidence cause his self-deception which in turn, cause his subjective well being which feeds back into his self-deception.

One might argue that the notion of cause is ambiguous or mysterious. We are suggesting a way to disambiguate or demystify the causal relations involved in self-deception by proposing that the mechanism of the causal relations is emotional coherence. Thus, the causal links in self-deception may be as depicted in Figure 4, but the way different causes lead to self-deception is through emotional coherence. The successful modeling of Dimmesdale's self-deception in HOTCO 2, an implementation of emotional coherence, supports this conclusion.

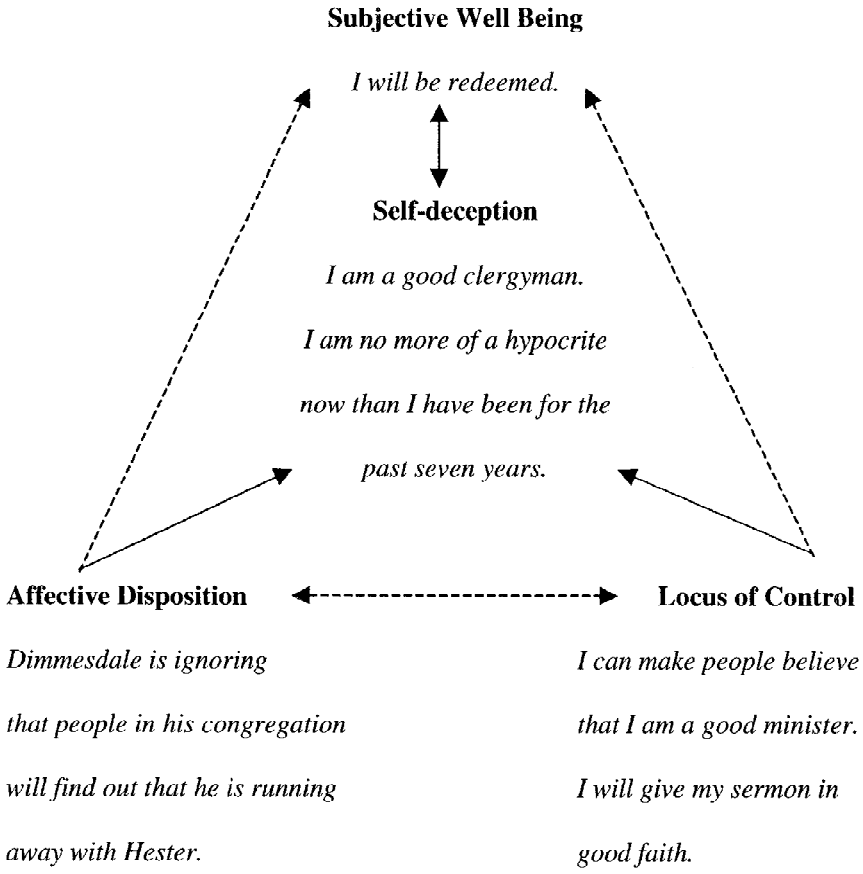

Figure 4. Causal Model of Dimmesdale's Self-deception. Adapted from Erez et al. (1995).

9. Wishful Thinking, Denial and Self-Deception

Wishful thinking is importantly different from self-deception. In wishful thinking an agent believes whatever he or she wants. Elster (1983), Mele (2001), and Johnston (1988) propose that at least some cases of self-deception can be explained in terms of wishful thinking. Although, it is important that the agent desires that p to self-deceive herself into believing that p, we think that self-deception is not just wishful thinking. In HOTCO 2, the valence input can be varied so as to make the system less or more emotional. After a certain degree of valence input, the system becomes so emotional that it models an agent who believes every single proposition that he or she deems as important. In a sense, emotions completely override reason. At such a degree of emotional input, the model shows that Dimmesdale not only believes that he is a good minister, he also believes that he will be redeemed. However, he does not think of himself as worthy of redemption in the experiments in which he successfully deceives himself into believing that he is a good clergyman.

Thus, in wishful thinking, people believe everything that they want to believe. Self-deception, however, is a 'weaker' state in that we may successfully deceive ourselves into believing something, but not everything that we wish to believe. This claim is supported by psychological studies on motivated inference (Kunda, 1999; Sanitioso et al., 1990); psychologists have shown that our judgments are colored by our motivations because the process of constructing justifications is biased by our goals. Nevertheless, we cannot believe whatever we want to believe. As Kunda (1999, p. 224) puts it, "Even when we are motivated to arrive at a particular conclusion, we are also motivated to be rational and to construct a justification for our desired conclusion that would persuade a dispassionate observer." There are constraints on what we can believe. In self-deception, we succeed in believing some (false) beliefs but not in believing everything we want to believe. Because some wishes remain unfulfilled, anxiety or internal conflict typically but not necessarily accompanies self-deception (as we will discuss in a coming section), but not wishful thinking.

Denial is also different from self-deception in that it is a kind of direct lie that self-deception is not. In denial, a person knowingly or consciously lies that $\sim p$. In self-deception, however, the person really *believes* that $\sim p$. Both claim something false, but in self-deception the correct belief (that p) is 'held' nonconsciously, whereas in denial, it is believed consciously. Also, in addition to denial, self-deception contains a very strong ego-enhancement component (Paulhus and Reid, 1991). Thus, self-deception and denial are importantly different.

10. Debates

In the beginning of our paper, we mentioned the debates over whether self-deception is intentional or not, and whether it involves a divided self or not. In this section we briefly comment on these debates. We also discuss the issue of whether the desire that p has to be "anxious" or not.

Regarding the issue of whether self-deception is intentional or not, we think that the debate is misplaced. To the extent that Dimmesdale intends to be redeemed, his self-deception can be seen as intentional. However, it would be absurd to claim that he intends to have the emotional preferences that he does. He may not have any control over his emotions at all. Emotional coherence occurs unconsciously. In their classic experiment pioneering the psychological studies of self-deception, Sackeim and Gur found that there were discrepancies in subjects' conscious misindentifications and nonconscious correct identifications (as indicated by Galvanic Skin Reponses) of voices as their own or others (Sackeim and Gur, 1979). They found that such discrepancies were a function of the subjects' independent manipulation of their self-esteem. (The subjects misidentified their own voices as others' when they thought ill of themselves, and others' voices as their own when they thought well of themselves.) We are proposing that the unconscious

mechanism behind self-deception is emotional coherence. The subjective goal of redemption, in Dimmesdale's case for instance, may be a conscious goal. However, the goal is approached through nonconscious emotional coherence. Is having the intent to be redeemed sufficient to call Dimmesdale's self-deception as intentional? No, for self-deception is not just approaching or avoiding one's goals. *How* the goal is approached or avoided is crucially a part of self-deception. One may fully intend to do whatever is necessary to achieve the desired goal, but at the same time, one does not *intend* to achieve emotional coherence involved in the approach of the goal. Therefore, until there is a good account of the relation between intentions and emotions, we cannot decide whether self-deception is intentional or not.

On the issue of a possible division of self involved in self-deception, we think that the issue arises from a misunderstanding of the notion of the self itself. Researchers have mainly focused on the deception side of self-deception, while rarely talking about the self of self-deception. What is the self that is deceived? In Dimmesdale's case, the narrator informs us of a tension between his "profounder self" and his presentational or outer self. However, this does not imply that there is necessarily a Freudian split in his self. It is not that Dimmesdale has two selves, self-A and self-B, and that self-A deceives self-B. There is a sense in which the self has a kind of continuity or oneness to it. However, the self is a "decentered, distributed, and multiplex" phenomenon that is "the sum total of its narratives, and includes within itself all the equivocations, contradictions, struggles and hidden messages that find expression in personal life" (Gallagher, 2000, p. 20). It is *because* the self is multiplex and devoid of any center, that self-deception is possible. Thus, if there is a 'split' in self, it is not at one place, but all over the place, and even in the self that does not deceive itself.

There is another debate worth paying attention to. Everybody agrees that in self-deception the agent, say, A is motivationally biased to believe p. Mele (2001) holds that the biasing role is played by A's desire that p. However, following Johnston (1988), Barnes (1997) insists that the desire that p must be anxious in that the person is uncertain of the truth of p. We can easily think of Dimmesdale's case as involving the desire to be redeemed. There is no doubt that if he wants to be redeemed. We can also say that he desires to be a good clergyman, and successfully deceives himself into believing that he is a good clergyman. Hawthorne makes it clear that Dimmesdale's self-deception causes him so much confusion that he experiences profound identity crises. It appears that in Dimmesdale's case his desire is anxious. However, we are inclined to agree with Mele that in self-deception the desire that p does not *have to* be an anxious desire.

There is good, although not conclusive computational evidence from our model that has inclined us to say that Mele is probably right on this issue. We conducted several hot experiments with varying degrees of emotional (valence) input in the system. The general trend was that the greater the valence input, that is, the more emotional the system, the easier (that is, faster) it was for the system to model self-deception. This suggests that the 'influence' of emotions on the system was a matter

of degree. The same may be true in humans. There is psychological evidence to suggest that self-deception is a dispositional tendency (Sackeim and Gur, 1979). It is plausible to hypothesize that this tendency is due to emotions and that depending on the degree to which people are emotional, they may be less or more disposed to deceive themselves.

11. Our Model Compared to Other Computational Models of Self-Deception

Rey (1988) gives a computational model based on the distinction between "central" and "avowed" attitudes of a self-deceiver. According to Rey, self-deception arises due to the discrepancies between the two kinds of attitudes. However, as Rey correctly notes, it is crucial that the discrepancies be motivated (p. 281). Otherwise, the agent would be self-ignorant and not self-deceiving. What is missing in Rey's model is any detailed account of what plays the motivated biasing role essential for self-deception. Our model shows that emotional coherence involving goals can provide the necessary motivated biasing.

Another notable model is Talbott's (1995) Bayesian model. Insofar as Talbott bases his model on the assumption of the self as a practically rational Bayesian agent, Talbott's model inherits the problems of a probabilistic approach to human thinking. The problems with probabilistic models of human thinking are discussed at length in Thagard (2000, Ch. 8). Such accounts assume that quantities that comply with the mathematical theory of probability can adequately describe the degrees of belief that people have in various propositions. However, there is much empirical evidence to show that human thinking is often not in accord with the notions of probability theory (see, e.g., Kahneman et al., 1982; Tversky and Koehler, 1994). On the other hand, as discussed in detail in Thagard (2000), coherence-based reasoning (on which our model is based) is pervasive in human thinking, in domains such as perception, decision making, ethical judgments, and emotion. Thus, our model is much more psychologically realistic than Talbott's model. In addition, Talbott fails to note the role of emotions in self-deception, whereas we have shown that emotions play a pivotal role in this phenomenon.

12. Conclusion

We have given a detailed analysis of a complex case of self-deception, namely, that of Dimmesdale in *The Scarlet Letter*. We have shown, by modeling Dimmesdale's self-deception in HOTCO 2, that self-deception can be seen as resulting from emotional coherence involving beliefs and goals. We have also compared our model to other models and have argued that our model is more psychologically realistic.

Appendix: Technical Details

The explanatory coherence program ECHO creates a network of units with explanatory and inhibitory links, then makes inferences by spreading activation through the network (Thagard, 1992). The activation of a unit j, a_j, is updated according to the following equation:

$a_j(t+1) = a_j(t)(1\text{-}d) + net_j(max - a_j(t))$
if $net_j > 0$, otherwise $net_j(a_j(t) - min)$.

Here d is a decay parameter (say 0.05) that decrements each unit at every cycle, *min* is a minimum activation (−1), *max* is maximum activation (1). Based on the weight w_{ij} between each unit i and j, we can calculate net_j, the net input to a unit, by:

$$net_j = \sum_i w_{ij} a_i(t). \quad (1)$$

In HOTCO, units have valences as well as activations. The valence of a unit u_j is the sum of the results of multiplying, for all units u_i to which it is linked, the activation of u_i times the valence of u_i, times the weight of the link between u_i and u_j. The actual equation used in HOTCO to update the valence v_j of unit$j$ is similar to the equation for updating activations::

$v_j(t+1) = v_j(t)(1\text{-}d) + net_j(max\text{-}\ v_j(t))$
if $net_j > 0$, $net_j(v_j(t) - min)$ *otherwise.*

Again d is a decay parameter (say 0.05) that decrements each unit at every cycle, *min* is a minimum valence (−1), *max* is maximum valence (1). Based on the weight w_{ij} between each unit i and j, we can calculate net_j, the net valence input to a unit, by:

$$net_j = \sum_i w_{ij} v_i(t) a_i(t).$$

Updating valences is just like updating activations plus the inclusion of a multiplicative factor for valences.

HOTCO 2 allows units to have their activations influenced by both input activations and input valences. The basic equation for updating activations is the same as the one given for ECHO above, but the net input is defined by a combination of activations and valences:

$$net_j = \sum_i w_{ij} a_i(t) + \sum_i w_{ij} v_i(t) a_i(t).$$

ECHO and HOTCO both proceed in two stages. First, input about explanatory and other relations generates a network of units and links. The LISP input for all simulations used in this paper is available on the Web at http://cogsci.uwaterloo.ca/coherencecode/co here/hotco-input.lisp.html. Second, activations and (for HOTCO)

valences are updated in parallel in accord with the above equations. Updating proceeds until all activations have reached stable values, which usually takes about 100 iterations of updating.

References

Arabsheibani, G., D. de Meza et al. (2000), 'And a Vision Appeared Unto Them of a Great Profit: Evidence of Self-Deception Among the Self-Employed,' *Economics Letters* 67, pp. 35–41.

Barnes, A. (1997), *Seeing Through Self-Deception*, Cambridge: Cambridge University Press.

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

de Sousa, R. B. (1988), 'Emotion and Self-Deception', in B. P. McLaughlin and A. O. Rorty, eds., *Perspectives on Self-Deception*, Berkeley, CA: University of California Press, pp. 325–341.

Demos, N. F. (1960), 'Lying to oneself,' *Journal of Philosophy* 57, pp. 588–595.

Elster, J. (1983), *Sour Grapes*, New York: Cambridge University Press.

Erez, A., D. E. Johnson et al. (1995), 'Self-Deception as a Mediator of the Relationship between Dispositions and Subjective Well-Being,' *Personality and Individual Differences* 19(5), pp. 597–612.

Fingarette, H. (1969), *Self-Deception*, London: Routledge and Kegan Paul.

Gallagher, S. (2000), 'Philosophical Conceptions of the Self: Implications for Cognitive Science,' *Trends in Cognitive Science* 4(1), pp. 14–21.

Goldbeck, R. (1997), 'Denial in Physical Illness,' *Journal of Psychosomatic Research* 43(6), pp. 575–593.

Goldberg, S. C. (1997), 'The Very Idea of Computer Self-Knowledge and Self-Deception,' *Minds and Machines* 7, pp. 515–529.

Haight, M. R. (1980), *A Study of Self-Deception*, Sussex: Harvester Press.

Harris, K. M. (1988), *Hypocrisy and Self-Deception in Hawthorne's Fiction*, Charlottesville, VA: University Press of Virginia.

Hawthorne, N. (1850), *The Scarlet Letter: A Romance*, Boston: Ticknor and Fields.

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

Kahneman, D., P. Slovic et al. (1982), *Judgment under uncertainty: Heuristics and biases*, New York: Cambridge University Press.

Kunda, Z. (1999), *Social Cognition*, Cambridge, MA: MIT Press.

Kipp, D. (1980), 'On self-deception,' *Philosophical Quarterly* 30, 305–317.

Lajunen, T., A. Corry et al. (1996), 'Impression Management and Self-Deception in Traffic Behavior Inventories,' *Personality and Individual Differences* 22(3), pp. 341–353.

Lazar, A. (1999), 'Deceiving Oneself Or Self-Deceived? On the Formation of Beliefs "Under the Influence,"' *Mind* 108(430), pp. 265–290.

McLaughlin, B. P. (1988), 'Exploring the Possibility of Self-Deception in Belief', in B. P. McLaughlin and A. O. Rorty, eds., *Perspectives on Self-Deception*, Berkeley: University of California Press, pp. 29–62.

McLaughlin, B. P. (1996), 'On the Very Possibility of Self-Deception', in R. T. Ames and W. Dissanayake, eds., *Self and Deception: A Cross-Cultural Philosophical Enquiry*, New York: SUNY.

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

Mitchell, J. (2000), 'Living a Lie: Self-Deception, Habit, and Social Roles,' *Human Studies* 23, pp. 145–156.

Paluch, S. (1967), 'Self-deception,' *Inquiry* 10, pp. 268–278.

Paulhus and Reid (1991), 'Enhancement and denial in social desirable responding,' *Journal of Personality and Social Psychology* 60, pp. 307–317.

Pears, D. (1982), 'Motivated Irrationality, Freudian Theory, and Cognitive Dissonance', in R. Wollheim and J. Hopkins, eds., *Philosophical Essays on Freud*, Cambridge: Cambridge University Press, pp. 264–288.

Pears, D. (1986), 'The Goals and Strategies of Self-Deception', in J. Elster, ed., *The Multiple Self*, Cambridge: Cambridge University Press, pp. 59–78.

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

Rorty, A. O. (1988), 'The Deceptive Self: Liars, Layers, and Lairs', in B. P. McLaughlin and A. O. Rorty, eds., *Perspectives on Self-Deception*, Berkeley: University of California Press, pp. 11–28.

Rorty, A. O. (1996), 'User-Friendly Self-Deception: a Traveler's Manual', in R. T. Ames and W. Dissanayake, eds., *Self and Deception: A Cross-Cultural Philosophical Enquiry*, New York: SUNY.

Sackeim, H. A. and R. C. Gur (1978), 'Self-Deception, Self-Confrontation, and Consciousness', in G. E. S. D. Shapiro, ed., *Consciousness and Self-regulation: Advances in Research*, New York: Plenum, pp. 139–197.

Sackeim, H. A. and R. C. Gur (1979), 'Self-Deception, Other Deception and Self-Reported Psychopathy,' *Journal of Consulting and Clinical Psychology* 47, pp. 213–215.

Sanitioso, R., Z. Kunda et al. (1990), 'Motivated Recruitment of Autobiographical Memories,' *Journal of Personality and Social Psychology* 59, pp. 229–241.

Sartre, J.-P. (1958), *Being and Nothingness*, London: Methuen.

Solomon, R. C. (1996), 'Self, Deception and Self-Deception in Philosophy', in R. T. Ames and W. Dissanayake, eds., *Self and Deception: A Cross-Cultural Philosophical Enquiry*, New York: SUNY.

Starek, J. E. and C. F. Keating (1991), 'Self-Deception and Its Relationship To Success in Competition,' *Basic and Applied Social Psychology* 12, pp. 145–155.

Surbey, M. K. and J. J. McNally (1997), 'Self-Deception as a Mediator of Cooperation and Defection in Varying Social Contexts Described in the Iterated Prisoner's Dilemma,' *Evolution and Human Behavior* 18(6), pp. 417–435.

Talbott, W. J. (1995), 'Intentional Self-Deception in a Single Coherent Self,' *Philosophy and Phenomelogical Research* LV(1), pp. 27–74.

Taylor, S. E. (1989), *Positive Illusions: Creative Self-Deception and the Healthy Mind*, New York: Basic Books.

Thagard, P. (1992), *Conceptual revolutions*, Princeton: Princeton University Press.

Thagard, P. (2000), *Coherence in thought and action*, Cambridge, MA: MIT Press.

Thagard, P. (forthcoming), 'Why Wasn't O. J. Convicted: Emotional Coherence in Legal Inference', *Cognition and Emotions*.

Thagard, P. and K. Verbeurgt (1998), 'Coherence as Constraint Satisfaction,' *Cognitive Science* 22, pp. 1–24.

Tversky, A. and D. J. Koehler (1994), 'Support Theory: A Nonextensional Representation of Subjective Probability,' *Psychological Review* 101, pp. 547–567.

Wrangham, R. (1999), 'Is Military Incompetence Adaptive?' *Evolution and Human Behavior* 20, pp. 3–17.

Zerbe, W. J. and D. L. Paulhus (1987), 'Socially Desirable Responding in Organized Behavior: A Reconception,' *Academy of Management Review* 12, pp. 250–264.