teorema
Vol. XXVI/3, 2007, pp. 131-143
ISSN: 0210-1602

Attributions of Self-Deception

Fernando Martínez Manrique

RESUMEN

El autoengaño es un tipo de estado mental que atribuimos habitualmente, producto de un sistema adaptativo y funcional para la protección del yo y la regulación de metas. Este artículo bosqueja un modelo para las funciones del auto-engaño, el "mecanismo de supresión", extrayendo consecuencias para los enfoques de la teoría y de la simulación en la atribución de estados mentales. Argumento que el "supresor" no presenta problema para la primera y delineo una teoría simple para la atribución de auto-engaño basada en las condiciones de accesibilidad y sinceridad. El supresor sí plantea una dificultad al enfoque de la simulación: las atribuciones de autoengaño sólo serían posibles si la simulación es defectuosa.

ABSTRACT

Self-deception is a kind of mental state that we ordinarily ascribe. It can be seen as the product of a functional system for protection of the self and regulation of goals. I sketch a model in which those functions may be realized, the suppression mechanism, so as to draw consequences for theory-based and simulation-based accounts of mindreading. I contend that the suppressor poses no problem for the former, outlining a simple theory-based account for the attribution of self-deception in terms of the conditions of accessibility and sincerity. However, the suppressor presents a difficulty for simulationism: self-deception would only be ascribed when the simulation is not perfect.

Self-deception is not an oddity of mental life. As Sahdra and Thagard have emphasized, "there is empirical evidence that self-deception is not only possible but also highly pervasive in human life" [Sahdra and Thagard (2003), p. 213]. Much effort has been devoted to analyze self-deception itself, with positions falling roughly into two different camps: those who see it as an intentional phenomenon, analogous to interpersonal deceit [e.g., Davidson (1985); Rorty (1988); Bermúdez (2000)], and those who regard it as non-intentional, typically fuelled by different sorts of motivational states [e.g., Johnston (1988); Barnes (1997); Mele (2001)]. Yet even though self-deception is a mental state that we indeed attribute — more often than one might expect — either to ourselves or to others, there is not so much work on how and

when we perform such attributions. Attribution of mental states is the province of mindreading, a field in which there are two competing theoretical positions: Theory Theory and Simulation Theory. Hence, in principle, two possible accounts for attributions of self-deception are possible.

In this paper I will examine the grounds for attributing self-deception, taking as an assumption that self-deception is an adaptive mechanism [Lockard and Paulhus (1988)], that seems to demand a specific functional characterization in the mind. Thus, I will first summarize the functions that self-deception may fulfil, and I will sketch a way in which those functions may be realized. Then I will outline theory-based and simulation-based accounts of attributions of self-deception. I will present a problem for the latter, arguing that, if it is correct, then self-deception can only be ascribed when the simulation is not perfect[1].

I. FUNCTIONS OF SELF-DECEPTION

Very roughly, cases of self-deception typically involve situations in which we would attribute possession of a belief to some individual, yet the individual sincerely disclaims having the belief in question. The adaptive value of self-deception can be summarized in three main functions:

> *Protect the self*. Self-deception can underlie psychological defences that preserve self-image, prevent harm to self-esteem, etc. [Nesse and Lloyd (1992); Paulhus and Suedfeld (1988)]. A classical example is that of an oncologist that disavows the belief that she has a tumour, despite having observed a number of symptoms that provide clear evidence to the contrary. Presumably the knowledge is too painful to be consciously accepted.
>
> *Achievement of goals.* When a belief interferes with a goal (e.g., showing some undesirable aspect of the goal), self-deception can play an instrumental role to achieve it by focusing attention on other, more acceptable, elements. In this case self-deception can be a variety of wishful thinking [Barnes (1997); Johnston (1988)]. An example of this is a case in which I wish to read the newspaper but I believe I should not do it during my working time. I focus my attention on the importance of being well-informed and form the belief that reading the paper is something really worth the time. After spending most of the morning doing so, I realize that the latter belief was the outcome of a self-deceiving manoeuvre in order to disregard the initial interfering belief.
>
> *Deceiving others*. Concealing one's own mental states from oneself may be an effective way to avoid "giving off" information in attempts

> at deception [Trivers (1985)]. For instance, Jones wants Smith's financial help for some investments, and Jones does not think that they would bring benefits to Smith. Yet, in the effort to be convincing, Jones persuades herself that it is really good for Smith. Rather than intentionally deceiving Smith, Jones is deceiving himself inasmuch as she sincerely holds the belief that Smith will benefit from the investment.

The three functions may be related. For instance, deceiving others can be seen as a particular case of achieving a goal in an interpersonal context; and it may also be the case that a belief interferes with a goal because it is harmful for self-esteem, so blocking the belief helps fulfilling both the protection of the self and the achievement of the goal. But they can also stand by themselves: a belief that threatens self-esteem can always be avoided by itself, not to achieve a subsequent goal; interfering beliefs do not need to threaten self-esteem; and achievement of goals does not need to involve the interpersonal component of deception. At any rate, what the three functions have in common is that they involve some *filtering* of information. Let me call this hypothetical functional system the *suppression mechanism* (or *suppressor*, for short).

II. THE SUPPRESSION MECHANISM

Self-deception cannot accomplish its functions if it works too overtly. If the oncologist says 'I know I have cancer, but I am going to dismiss this belief', it is unlikely that she will be able to protect herself from the influence of the painful belief. Likewise, if Jones consciously thinks 'I am going to convince myself that the investment is beneficial for Smith', Jones can be regarded as cynical, not self-deceptive. It has been suggested [Greenwald (1988); Paulhus and Suedfeld (1988)] that some sort of filtering mechanism underlies self-deception. Along this line, I will hypothesize that self-deception relies on some processing mechanism whose job is to suppress selectively beliefs. We can call this mechanism the *suppressor*. To suppress a belief is to make it inaccessible to consciousness. To function properly, the suppressing process itself should be opaque to conscious access. Hence, the suppressor cannot be under voluntary control (unlike other processes, like memory searching or attention focusing), and the subject can never be sure that the suppressor is acting. To keep things simple I will regard the suppressor as a single functional box, even though the mechanism that is actually in charge of this task has to be far more complex.

In order to see what a suppressor would do it is useful to have a model of the access to our own mental states. I will take for my purposes the Monitoring Mechanism (MM) proposed by Nichols and Stich (2003). MM is a distinct mechanism specialized in the detection of one's propositional attitudes.

For instance, assuming the existence of a "belief box" for all the states functionally characterized as beliefs, MM takes the tacit belief P and produces an explicit belief in the subject about his possession of belief P. The subject is then aware of this belief. The suppressor must work somewhere before MM outputs its representations. If self-deception fulfils the three kinds of functions mentioned above, the suppressor must be "in touch" with at least two other systems: (a) an evaluative system that judges if the belief is potentially dangerous for the self, (b) a practical reasoning system that establishes whether the belief can move the subject closer to a given goal. Emotion is probably part of such an evaluative system, while the subject's goals can be functionally condensed in a "desire box".

Figure 1 shows a sketch of the model. The belief P is tacitly held in the belief box. It is picked up by a mechanism that monitors the box, but it reaches awareness only if the suppressor does not block it. The suppressor "vetoes" the entry when (a) the belief would conflict with some relevant goal of the subject as determined by the practical reasoning system, or (b) the belief would potentially induce some aversive state (e.g., a decrease in self-esteem), according to the judgment of an emotional evaluative system.

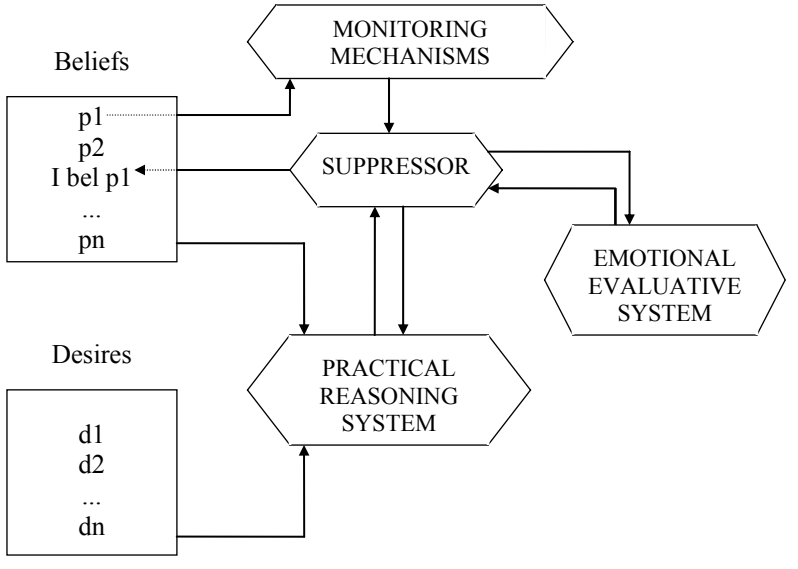

Figure 1. Suppressor mechanism (adapted from Nichols and Stich, 2003)

Let me stress that the idea of a suppressor is not dependent on the MM model. It can be incorporated to other theories that hypothesize some mechanism(s) that "mark" information as conscious. For instance, in Carruther's Reflexive Thinking theory [Carruthers (1996)] a suppressor might be hypothesized as a filter of information from the Reflexive Thinking component to the Conscious Short-Term Store. Of course, there are ways to consciousness that involve neither the Monitoring Mechanism nor the Reflexive Thinking component, e.g., both Nichols and Stich, and Carruthers include perceptual paths to consciousness. A suppressor would not operate on those paths, presumably because it would not be adaptive to do so.

There are many questions that an account like this must face. Even though I offer it mainly for heuristic purposes, let me address three possible objections. First, we can wonder why a belief that leads to an aversive emotional response should be suppressed, when these kinds of responses are clearly adaptive, i.e., they may lead to courses of action to stop or avoid the aversive situation. A possible answer comes by means of an analogy between the suppressor effect and the effect of self-created analgesics (e.g., endorphins) that the organism produces to control pain [Nesse and Lloyd (1992)]. Self-deception would act as a way to make the situation tolerable and, by deviating attention from the damaging belief, it would even allow the individual to devote cognitive resources to other goals that can get frustrated by excessive attention to the painful belief.

Second, it is central to standard theories of cognition that many beliefs can control behaviour without becoming conscious at all. Then it is not clear how the suppressor could prevent a belief from interfering with other goals. The answer presumably has to do with the greater activation that conscious beliefs get. The key idea is that beliefs (and desires) compete with each other for control. Suppression biases this competition by cutting the access of certain beliefs to consciousness, and hence to cognitive resources like attention or memory. If these beliefs are still powerful enough they may keep influencing behaviour unconsciously, especially when the alternative "favoured" beliefs are weak. This may be particularly the case when suppression is oriented to protect the self, with no further goals to achieve. We would have a type of self-deception in which a belief is verbally disclaimed but some other system shows to be under its control, perhaps an automatic motor system (e.g., a person who affirms that he is not afraid, but whose response is to flee), or a complex of belief-involving mechanisms (e.g., the oncologist of the previous example, who suddenly begins to send affectionate letters to old friends and relatives).

Third, there are cases in which one *actively* makes oneself reject an unwanted belief. The belief became conscious, yet the subject managed to expurgate it anyway. This seems to be a case in which self-deception is (partially at least) under voluntary control. Now, the suppressor is an automatic mechanism. So we might think that there are other ways to self-deception apart

from the suppressor. A possible answer is that even if these other ways might exist, they are not really necessary. It can be argued that what the subject does in active self-deception is to mobilize alternative goals and beliefs, to make them especially salient and operative, so that the suppressor gets finally triggered. This is similar to those cases in which someone self-activates her autonomic nervous system, e.g., bringing about thoughts that make her furious. It is the thinking that is voluntary, not the response itself.

Having sketched a model of functional self-deception, I turn to examine attributions of self-deception. I will show that if the model is right, it does not affect theory-based accounts of attribution of self-deception, i.e., the model seems to fit nicely with the assumptions underlying those accounts. However, the effects on simulation-based accounts are more serious: I will argue that *if* there is a filtering mechanism of the kind just presented, then simulationism cannot provide a successful account of attributions of self-deception.

III. Attributing self-deception in the Theory Theory

According to the Theory Theory (henceforth TT), when we predict/ explain someone's behaviour we use a body of (mostly tacit) theoretical information about folk psychology. Erroneous explanations are often due to some deficiency in this data base, i.e., the lack of a relevant piece of information or the possession of an incorrect one. There are different ways in which TT can be construed [see Davies and Stone (1995a), (1995b); Gopnik and Meltzoff (1997)] but we can assume that the shape of attributions of self-deception is roughly the same in them.

Obviously, the idea that minds have suppressors (or, for that matter, any other processing mechanism) does not need to belong to the theory that people use to make attributions of mental states. But it is the case that people *do* sometimes conclude that someone is self-deceiving. When do we say 'she is deceiving herself', rather than 'she is wrong' or 'she is cheating'? Typical cases seem to involve conditions in which the truth is "at hand" for the subject, while she sincerely disclaims acquaintance of it. Let me outline a simple theory of attributions of self-deception: we judge that a person is self-deceiving about P when (i) we would normally attribute to her the belief that P, (ii) she disavows the belief that P, and (iii) the following two conditions apply:

(iii.1) *Accessibility condition*:
P is (potentially) highly accessible for the subject[2].

(iii.2) *Sincerity condition*:
The subject is sincere in her disavowal of belief P.

The theory is simple because it relies on two conditions that allow us to distinguish self-deception from other deception-related phenomena that are sometimes seen as akin to it [e.g., Rey (1988)]. Figure 2 summarizes these distinctions.

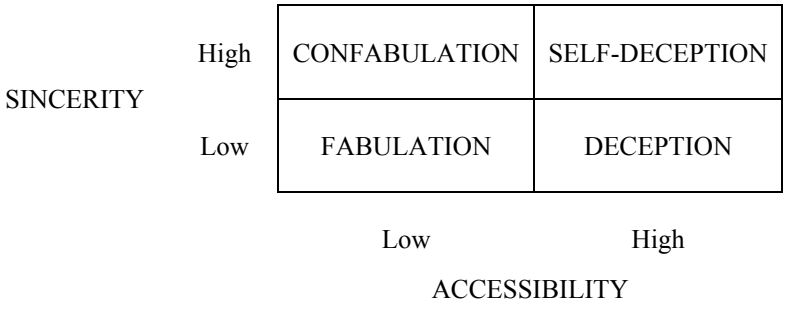

Figure 2. Dimensions of deception-attribution to others

When we think that belief P is accessible to the subject but we have grounds to doubt about her sincerity, we have a case of *deception*: we judge that the person is simply lying. In contrast, when we do not doubt about the sincerity of the subject but we think that P is not easily accessible to her, we can regard the case as one of *confabulation* [Nisbett and Wilson (1977)]. In cases of confabulation people give reasons that have nothing to do with the actual mental states that explain their behaviour. In a classical example of this, the subject's choice of some garment is demonstrably influenced by its relative position to the subject (e.g., subjects tend to choose the one to the right). However, the subject claims that the position did not affect her choice. Instead she believes that another factor (e.g., quality of fabric) was behind it. We admit that the subject is sincere but we do not count this as a case of self-deception. The belief that this judgment is caused by the position of the garment does not seem to figure among the beliefs that subjects are normally able to report. (This is reflected by the fact that none of the experimental subjects mentioned position among the factors for their choice, even if the position demonstrably had non-random influence in the vast majority of them).

When neither the accessibility nor the sincerity conditions are met, we have what we may call *fabulation*. Take the experiment with judgments about garments again, but suppose this time that the subject does not believe her own explanation. Although she knows that the quality of the garment has nothing to do with her choice, she does not know what motivated it. How-

ever, instead of confabulating her explanation, she deliberately concocts one, knowing that it is pure invention. This is a case of mild deception: the subject is lying, not because she is concealing the truth, but because the truth is also unknown to her. It might be that many cases of lying in young children belong to this category: a poor understanding of her own mental states leads the child to fabulate alternative explanations that she does not take "too seriously".

Rey (1988) offers a similar computational account in which self-deception arises as a conflict between central and avowed beliefs. An agent is self-deceived when she centrally holds P, but preferring not to avow it she prevents herself from doing so. From the point of view of the observer that concerns us here, the question is when it is possible to conclude that central beliefs differ in the appropriate way from avowed beliefs. To infer the subject's central beliefs we may rely either on her behaviour or on her sources of information. If her behaviour is incongruent with her avowed beliefs (and, as I explained above, this can be especially the case when self-deception fulfils the function of protecting the self) we may take this as an indication of self-deception. If the behavior is congruent (and this can be especially the case when self-deception fulfils functions related to achievement of goals) we may still look for evidence of self-deception in the sources of knowledge available to her, in her past avowals of beliefs, and so on.

Sincerity and accessibility conditions can thus reflect different ways to gather evidence in order to explain someone's behaviour[3]. On the one hand, there is some support for the existence of specific mechanisms devoted to the detection of deceit [Cosmides and Tooby (1992); Ekman (1988)]. On the other one, evidence of the information accessible to the subject can be obtained from means that range from mere observation of her perceptual surrounding to detailed interrogation to probe whether some element is lacking, she neglected something, etc. The pieces of knowledge gained by these different means would be combined in an explanatory theory. If both sincerity checking and information-accessibility checking procedures point toward the "high" end we are more likely to conclude that the individual is self-deceptive.

IV. Attributing self-deception in the Simulation Theory

According to the Simulation Theory (henceforth ST) when we attribute mental states to a subject we position ourselves imaginatively in the subject's condition and record the states that this condition would produce. Stich and Nichols (1992) characterized simulation as a process in which the predicting/explaining person lets her own reasoning processes run off-line (i.e., without bringing about an actual behaviour) and notes down the result. The important factor to produce a correct explanation is to feed these processes with the correct pretend inputs, that is, with inputs that correspond in the

most approximate way to those received by the subject whose behaviour is to be explained. Off-line simulation can be construed in more than one way [Davies and Stone (1995a), (1995b); Goldie (1999); Nichols *et al.*, (1996)], but I will ignore the differences for the limited purposes of this paper.

Like theory-based accounts, ST does not presuppose any knowledge of a suppressor in the attributing subjects. (In fact, ST does not need to suppose *any* folk psychological knowledge at all). However, ST needs to assume that the attributing subject has roughly the same mental apparatus as the subject to be attributed. Thus if the suppressor hypothesis is correct, both subjects have a suppressing mechanism (in normal conditions). But now there is a paradox for the simulation-based account of self-deception. Suppose that you run a simulation of a subject who is, in fact, self-deceiving. Suppose that your pretend inputs are absolutely accurate. Thus all the mechanisms that get activated in the simulated subject are also active in your simulation. One of these mechanisms is the suppressor. The upshot is that if the belief P was suppressed in the subject, *it will also be suppressed in you*. But if this is so, you will never come to conclude that the subject is self-deceiving about P. The suppressed belief P will never figure in your prediction/explanation, because it was concealed from yourself too. In its place, you will obtain whatever belief replaces P in the simulated subject's mind (i.e., the belief she is actually aware of). In other words, a perfect simulation may correctly attribute that someone *disclaims* P, but it cannot attribute that someone *self-deceives* about P.

Note that nothing like this happens in the Theory Theory: if we have a perfect theory we can reach the conclusion that a subject is self-deceptive about P. This is so because TT only demands the use of our reasoning capacities (possibly, both general and Theory-of-Mind-specific) operating on our sources of knowledge. The suppressor does not need to become activated in this process (and if it does get activated it will be for altogether different reasons). In contrast, a requisite of a good simulation is that the simulating person employs the same processes as the simulated one. As the suppressor is an automatic system, it will be triggered if the simulated inputs are sufficiently close to the real ones. Thus the simulating person will become self-deceived himself.

Several rejoinders are possible. One is to point out that perfect simulations rarely occur. So in practice the predictor can reach the suppressed belief and attribute it to the predicted subject. But this still would make simulation an odd process, at least in the case of attributing self-deception: it only succeeds when it gets something wrong. A second reply is to grant that simulation-based accounts only work for cases of prediction, not explanation, of behaviour, and that the alleged paradox would only arise in cases of prediction. Now, it is possible to make a case that self-deception is very seldom predicted. We do not usually conclude that someone *will* self-deceive; rather, we explain someone's behaviour saying that she *has* deceived herself. Attributing self-deception, one may say, comes as a (relative) surprise. Neverthe-

less, even if predictions of self-deception are rare, a simulationist account would make them impossible. A TT account, however, would allow prediction of self-deception in principle, conceding that prediction is unlikely because our theories are not usually powerful enough.

A third possible reply from ST is that attributions often involve several sequential steps. The simulating subject progressively adjusts the relevant information that she receives. For instance, here is a rough idea of how a simulation could run in a case like the self-deceived oncologist. First I launch a simulation of the oncologist facing the evidence of the positive tests, and conclude that she will believe that she has a tumour. Then I find out that she disclaims this belief. This is a new pretend input that I have to adjust, so I launch a second simulation in which I pretend to disclaim sincerely this belief. As the first and second simulation conflict, I conclude that the oncologist is self-deceiving about the belief previously obtained. However, even if something like this happens in simulation, the paradox does not vanish. It looks like we always need a previous failure in the simulation process to succeed in the explanation. Furthermore, it is noticeable that TT does not in principle require such a sequential procedure: we can combine concurrently all the relevant pieces for our explanatory theory.

The problem I am presenting for ST does not appear in other contexts. For instance, if a perfect simulation puts me in the same emotional state as the simulated individual, I can accurately determine which emotion it is. The problem does not arise either in the simulation of abnormal states (i.e., explaining the states of a schizophrenic). In these cases it is perfectly safe for the simulation theorist to suppose that some mechanism in the simulated person does not function in the same way as in the simulating one. But if self-deception is an adaptive system, then it is *not* abnormal. It is something that happens in ordinary circumstances, as anyone's experience of it reveals. If self-deception presents a specific problem for ST, I suggest, is because of the complex nature of the phenomenon. It is possible that simulation is always insufficient for complex attributions of mental states, like those that involve second-order beliefs. These cases would demand the arrangement of all the evidence in the form of a theory. Hence, if attributions of self-deception present a paradox for simulationist accounts, this should not necessarily imply an outright rejection of ST, only a limitation of the contexts in which simulation may take place.

V. Conclusion

I began this paper by endorsing two assumptions: first, that self-deception is a widespread phenomenon of mental life — indeed, a kind of mental state that we ordinarily ascribe; second, that self-deception is not an anomalous

condition but the product of an adaptive, functional system, more specifically, one that is involved in the protection of the self and the regulation of goals. Following these assumptions, I posited the presence of some automatic filtering mechanism that suppresses selectively the contents accepted in awareness, and then I drew the consequences for theories that deal with the attribution of mental states. The existence of such a mechanism does not pose a problem for Theory Theory accounts of how people attribute states of self-deception. In fact, I provided the outline of a simple theory-based account that distinguishes them from other related phenomena in terms of the accessibility of the disclaimed information, and the sincerity credited to the individual. On the other hand, the filtering mechanism gives rise to a paradox for simulationist accounts: a perfect simulation would not yield the outcome that someone is self-deceiving, since the simulating person would self-deceive himself. The moral is that the complexity of mental states and the mechanisms that instantiate them may affect the theories of how we attribute such states. If self-deception is actually a distinctive kind of mental state we'd better understand how it relates to the rest of mental life in order to know which theory of mindreading explains best the way we grasp it in our ordinary affairs.

Departamento de Filosofía
Facultad de Filosofía y Letras
Universidad de Granada
Campus de Cartuja, E-18071, Granada
E-mail: fmmanriq@ugr.es

NOTES

[1] I will have to make a number of simplifying assumptions: I will deal only with self-deception about beliefs, even though people may deceive themselves about other mental states, like emotions [de Sousa (1988); Griffiths (1997), pp. 150-55]; and I will treat all the beliefs as belonging to a single functional system — the "belief box" — even if there could be distinct belief systems that would demand specific self-deception mechanisms.

[2] The accessibility condition is related to (i) in the sense that if P does not seem accessible for the subject, then it is unlikely that we attribute her the belief that P. But (i) is not redundant: it is necessary to keep them separate because in some cases one may attribute the belief that P even with low accessibility, simply because it is needed to explain a particular behaviour — e.g., in the cases of confabulation explained below.

[3] I must remark that these conditions do not intend to provide an analysis of self-deception in terms of necessary and sufficient conditions for someone to be self-deceived, much less an analysis of the reasons or causes leading to a state of self-deceit. I only mean to outline the circumstances that favor a judgment of self-deception.

[4] This paper is part of research project HUM2005-07358 of the Ministerio de Educación y Ciencia. I want to thank Luc Faucher, as well as audiences at the Euro-

pean Society for Philosophy and Psychology where an earlier version of this paper was presented.

References

BARKOW, J., COSMIDES, L. and TOOBY, J. (eds.) (1992), *The Adapted Mind*, New York, Oxford University Press.

BARNES, A. (1997), *Seeing through Self-Deception,* Cambridge, Cambridge University Press.

BERMÚDEZ, J.L. (2000), "Self-Deception, Intentions, and Contradictory Beliefs", in *Analysis,* 60, pp. 309-319.

CARRUTHERS, P. (1996), *Language, Thought and Consciousness*, Cambridge, Cambridge University Press.

COSMIDES, L. and TOOBY, J. (1992), "Cognitive Adaptations for Social Exchange", in Barkow *et al.*, pp. 163-228.

DAVIDSON, D. (1985) "Deception and Division", in E. LePore and B. McLaughlin, (eds.), *Actions and Events* Oxford, Basil Blackwell, pp. 138-148.

DAVIES, M. and STONE, T. (eds.) (1995a), *Folk Psychology*, Oxford, Blackwell.

— (1995b), *Mental Simulation*, Oxford, Blackwell.

DE SOUSA, R. B. (1988), "Emotion and Self-Deception", in McLaughlin and Rorty, pp. 324-341.

EKMAN, P. (1988), "Self-Deception and the Detection of Misinformation", in Lockard and Paulhus, pp. 229-250.

GOLDIE, P. (1999), "How We Think of Others' Emotions", in *Mind and Language*, 14, pp. 394-423.

GOPNIK, A. and MELTZOFF, A. (1997), *Words, Thoughts, and Theories*, Cambridge, MA, MIT Press.

GREENWALD, A.G. (1988), "Self-Knowledge and Self-Deception", in Lockard and Paulhus, pp. 113-131.

GRIFFITHS, P.E. (1997), *What Emotions Really Are* Chicago, Chicago University Press.

JOHNSTON, M. (1988), "Self-Deception and the Nature of Mind", in McLaughlin and Rorty, pp. 63-91.

LOCKARD, J.S. and PAULHUS, D.L. (eds.) (1988), *Self-Deception: An Adaptive Mechanism?*, Englewood Cliffs, NJ, Prentice-Hall.

MCLAUGHLIN, B.P. and RORTY, A.O. (eds.) (1988), *Perspectives on Self-Deception* Berkeley, CA, University of California Press.

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

NESSE, R.M. and LLOYD, A.T. (1992), "The Evolution of Psychodynamic Mechanisms", in Barkow *et al.*, pp. 601-624.

NICHOLS, S. and STICH, S. (2003), "Reading One's Own Mind: A Cognitive Theory of Self-Awareness" in Q. Smith and A. Jokic (eds.), *Aspects of Consciousness*, Oxford, Oxford University Press, pp. 157-200.

NICHOLS, S., STICH, S., LESLIE, A. and KLEIN, D. (1996), "Varieties of Off-Line Simulation", in P. Carruthers and P.K. Smith (eds.), *Theories of Theories of Mind*, Cambridge, Cambridge University Press, pp. 39-74.

NISBETT, R. and WILSON, T. (1977), "On Saying More than We Can Know", in *Psychological Review*, 84, pp. 231-259.

PAULHUS, D.L. and SUEDFELD, P. (1988), "A Dynamic Complexity Model of Self-Deception", in Lockard and Paulhus, pp. 132-145.

REY, G. (1988), "Toward a Computational Account of *Akrasia* and Self-Deception", in McLaughlin and Rorty, pp. 264-296.

RORTY, A.O. (1988), "The Self-Deceptive Self: Liars, Layers and Lairs", in McLaughlin and Rorty, pp. 11-28.

SAHDRA, B. and THAGARD, P., (2003), "Self-Deception and Emotional Coherence", in *Minds and Machines*, 13, pp. 213-231.

STICH, S. and NICHOLS, S. (1992), "Folk Psychology: Simulation or Tacit Theory", reprinted in Davies and Stone (1995a), pp. 123-158.

TRIVERS, R.L. (1985), "Deceit and Self-Deception", in *Social Evolution*, Menlo Park, CA, Benjamin/Cummings, pp. 395-420.