CRÍTICA, Revista Hispanoamericana de Filosofía. Vol. 33, No. 97 (abril 2001): 3–32

MOTIVATED IRRATIONALITY: THE CASE OF SELF-DECEPTION*

MONTSERRAT BORDES
Facultad de Humanidades
Universidad Pompeu Fabra
montserrat.bordes@huma.upf.es

SUMMARY: This paper inquires into the conceptual nature of self-deception. I shall afford a theory which links SD to wishful thinking. First I present two rival models for the analysis of SD, and suggest reasons why the interpersonal model is flawed. It is necessary for supporters of this model to work out a strategy that avoids the ascription of inconsistency to the self-deceiver in order to fulfill the requirements of the charity principle. Some objections to the compartmentalization strategy are put forward, and a motivational theory is advanced. This theory diverges from Mele (1997)'s account of SD in that it (i) establishes as a necessary condition for SD the existence of a causal link between a desire and a belief unacknowledged by the self-deceived subject, who is unaware also of the counterevidential nature of his belief (the 'focused inferential blindness' thesis), (ii) allows only 'weak SD' cases and offers methodological reasons against the seemingly intentional and dissociative nature of SD and (iii) stresses the deception-SD asymmetry.

KEY WORDS: motivated irrationality, wishful thinking, self-deception, inconsistency

RESUMEN: El presente artículo intenta investigar la naturaleza conceptual del autoengaño. Presentaré dos modelos rivales de análisis y ofreceré razones contra la teoría interpersonal frente a la motivacional, alegando las dificultades que comporta seguir alguna de sus estrategias de compartimentación para evitar la atribución de inconsistencia simple al sujeto autoengañado. Defenderé una teoría que vincula el autoengaño con la creencia desiderativa. Se trata de una teoría motivacional que difiere de la de Mele (1997) en que (i) establece como condición necesaria para el autoengaño que se dé una relación causal entre el deseo y la creencia pertinentes, relación cuya existencia desconoce el sujeto, que ignora también el grado en que su creencia es incompatible con los datos empíricamente disponibles (tesis de la ceguera inferencial focalizada), (ii) acepta sólo casos de autoengaño débil y ofrece razones metodológicas contra la supuesta naturaleza intencional y disociativa del autoengaño y (iii) subraya la asimetría autoengaño/engaño.

PALABRAS CLAVE: irracionalidad motivada, creencia desiderativa, autoengaño, inconsistencia

[*] I am grateful for comments on a predecessor of this paper from Stephen Jakobson. An early draft of this paper has been presented at Warwick, ESPP'99.

In *De profundis* Oscar Wilde catholically expressed his sadness at the vitiated nature of his lover, whose mischief eventually led to the writer's imprisonment. Despite his lucid description of Douglas' egoistic behavior towards him, and given the dispositional nature of the emotions involved, Wilde's sincere claim that Lord Alfred really loved him may be seen as manifesting a state of self-deception (SD).

This paper is an attempt to inquire into the conceptual nature of self-deception. I shall afford a theory which links SD to wishful thinking and claims that SD is a self-centered mental state, an irrational doxastic disposition which consists in the conservative and unpurposeful holding of a belief against the weight of the available evidence. It is epistemically (not always instrumentally) irrational because it is desire-grounded, but, it is not related to the relevant evidence in the same way as wishful thinking is. I present two rival models for the analysis of SD, and suggest reasons why the interpersonal model is flawed. Because interpersonal theories characterize SD cases as intentional and dissociative ('strong SD'), it is necessary to work out a strategy avoiding the ascription of inconsistency to the self-deceiver in order to fulfill the requirements of the charity principle. Some objections to the compartmentalization strategy are put forward, and a motivational theory —more respectful of the charity principle and the exclusivity principle of belief— is advanced. This theory diverges from Mele (1997)'s motivational account of SD in that it (i) establishes as a necessary condition for SD the existence of a causal link between a desire and a belief unacknowledged by the self-deceived, unaware also of the counterevidential nature of his belief (the focused inferential blindness' thesis), (ii) allows only 'weak SD' cases and offers methodological reasons against the seemingly intentional and dissociative nature of SD, (iii) stresses the deception-SD asymmetry by claiming that the latter, not the former, is essentially involved with epistemic reliability and not with truth value, and (iv) is able to distinguish between SD and mistake due to ignorance.

Even if I reject Williams (1973) and (1978)'s thesis of the involuntariness of belief for Elsterian reasons, I take it that, because a desire has to be the proximal pervasive cause of the self-deceptive belief, SD is not acquired as a by-product of an indirect intentional strategy.

I will proceed as follows. First of all the main tenets of the interpersonal model of SD are examined in some depth and the notion of strong SD is presented. I then outline some of the interpersonal strategies to avoid the attribution of inconsistency to the subject and some of the objections to them. Thirdly I present my inferential blindness model of SD, which supplements Mele's motivational analysis with an inferential blindness thesis and a causal link condition, and offer some grounds for its plausibility. Finally I sum up the main requirements of my approach.

1. *The Interpersonal Theory of SD*

Like cases of akrasia, wishful thinking and adaptive preference cases (e.g., the sour grapes phenomena),[1] SD is the outcome of an illegitimate epistemological process of belief and/or decision-formation, a kind of 'cognitive dissonance', in Festinger (1957)'s terms.[2] All these modes of irrationality are motivated. I do not want to enter into the dispute about the concept of motivation. I will simply explain how I will conceive of it here. By a motive I understand a mental state or inclination whose occurrence precedes and partially causes an action or another mental state. A motive can be a desire (the fervent desire of having a short poem published in a literary journal of repute may motivate a counterevidential belief in the author that it has the required quality) or an emotion (fear facing a situation perceived as dangerous triggers flight). A motive is not an intention. Someone may act out of a motive, but do so unintentionally. As Kenny (1963, p. 87) notes, it is possible to act by a motive even if

[1] Its paradigm is the famous fable by Aesop —later popularized by La Fontaine— of the fox and the grapes, despised by the agent after realizing they were unattainable; an example of desire and behavior modification by conformist adaptation to the world.

[2] SD and akrasia can be seen respectively as types of belief/action formed against the best judgement the thinker/agent is able to come to. SD could be seen as a kind of akratic belief.

the agent does not have the concept of it; however, without the concept of the intention an agent cannot act by it.

SD cases are also, like the other phenomena referred to, irrational states. But it is epistemic irrationality, not instrumental irrationality which is at issue here. The self-deceived forms a belief that is not well-grounded, not evidentially warranted, so that he attempts against a rationality principle like Price (1965, p. 131)'s: the degree of assent to a proposition has to be proportional to the strength of the available evidence about it. The self-deceived, however, gives assent to a clearly counterevidential proposition. Now there is another sense in which a state or action is said to be irrational, for example, when it is not instrumentally able to attain a desired objective.[3] In this sense of the word, SD cannot definitionally be dubbed 'irrational', insofar as plenty of SD cases are benefic: they possess a high degree of instrumental value as adaptive mechanisms, optimizers of the mental equilibrium and anxiety reducers. The apprehensive oncologist's SD about his own cancer,[4] as acknowledgement of his disease would undermine his disposition to recover, is beneficial not only for his mental, but for his physical wealth too.

La Rochefoucauld, Montaigne and Pascal, all of whom were interested in the study of the etiology and functional relations between mental human events, alerted us to the generalized use of rationalization strategies for justifying our actions and aiming to conceal the true motives inspiring them (think, for example, of the doctor believing he chooses his profession impelled by altruistic purposes and not, as was really the case, by the social prestige attached to it). Some authors even defend the universality of SD as a feature always present in every human mind.[5] On

[3] On this see Davidson (1980)'s essays.

[4] The example comes from Rorty (1972, p. 401).

[5] See Sartre (1943, part i, chap. II) and La Rochefoucauld's maxims related to his 'mental veil' theory (as a result of our custom of concealing our mind from the other people, we conceal it from ourselves; see Elster (1999)'s comments) and how our self-esteem is related to SD phenomena.

the opposite side, sceptics deny its occurrence, alleging its inconsistency, which thus makes it an unascribable mental state.[6]

In this paper I shall limit myself to examining the moderate theories which accept the occasional exemplification of SD. I will ignore universalist and sceptical theories. Despite the fact that Freud did his best to obtain cultural citizenship for SD, even the most reluctant person to accept Freudian ontological extravagances may approach this notion from a minimalist deflationary framework, without being committed to concepts like "ego's disease", "authenticity" or "engagement with the world", appealed to by Kierkegaard, Sartre or even Fingarette, seemingly satisfied with an *obscurum per obscuris*.

The main moderate theorists embrace either the interpersonal model, or the motivational model. Those resorting to the first model analyse SD as a special case of deception.[7] For them 'X deceives himself' has to be read after the scheme 'X deceives Y' (interpersonal deception), so that SD is no more than the reflexive version of deception: 'X deceives X' (intrapersonal deception). Necessary and sufficient conditions for deception may be specified quite uncontentiously, as follows. Given two rational beings, X and Y, and a proposition, p, X deceives Y at time t' ($t>t'$) about that p if and only if:

(1) at t X intends to induce in Y the belief that p.
(2) at t' X succeeds in inducing in Y the belief that p.
(3) at t and t' X knows (or truly believes) that not-p.

This analysis entails that deception is deliberate, intentional (by (1)),[8] successful (by (2)) and that its object is a false proposi-

[6] Haight (1980).

[7] Among supporters of the interpersonal model are Demos (1960), Fingarette (1969), Rorty (1972), Pears (1975), Audi (1985), Davidson (1982) and (1985) and Guttenplan (1994).

[8] There is no focus of disagreement on these requirements except for the alleged necessity of (1). Mele (1983) admits cases of non intentional SD (against Siegler (1968), Demos (1960), Fingarette (1969), and many others). I will not address the occasional current use of the term in that sense and take note of the dictionary entry that to deceive someone is "to make believe someone something that is not true". Dictionaries define self-deception as not

tion (by (3)). (3) accounts for the fact that, in order to deceive Y, it is not necessary that X has at his disposal a correct grounding for not-p or that he has acquired the belief by a causal reliable epistemological process. It suffices with the truth of not-p and X's believing so.

A consequence that follows from this analysis applied to the reflexive case is that the overcoming of SD does not imply that X (as a deceived person) comes to know something of which he was previously ignorant, because he (as a deceiver) already knew that p. SD, as seen from the interpersonal model point of view, is not basically a kind of ignorance,[9] but a kind of implicit knowledge.

According to this model, SD is held to show the following features: (i) it is deliberate or intentioned, the outcome of intending to believe that not-p (SD's intentionality thesis) and (ii) it is paradoxical, as it consists in the possessing of incompatible beliefs. From (1)-(3) it follows that at t' X believes that p and that not-p (SD's conflicting belief thesis).

The intentionality thesis ascribes to SD an intentional feature and the conflicting belief thesis a dissociative one. I shall call SD cases showing both intentional and dissociative features 'strong

admitting to oneself something one knows to be true, but this should not persuade anyone to afford the interpersonal model for analysing SD! I take it that dictionaries have to address the equivocity of 'deception', related at least to three concepts: involuntary induction to mistake, mistake, and deception as such, which, unlike the other two, involves responsibility and intentionality. In any case, the current literature on SD has captured the concept basically in the latter sense.

[9] It is on the interpersonal model that the Wittgensteinian-minded philosopher has to rely in order to fit SD cases in the frame of his first-/third person asymmetry thesis. According to Hacker (1996)'s reading of it, the thesis implies that negative psychological first-person sentences (of the type 'I don't know whether F(I)', where 'F' is a mental predicate as 'feel pain', 'believe that p', 'desire that p',. . .) never describe ignorance states. They are either absurd (as 'I don't know whether I feel pain') or avowals of indecision (as 'I don't know whether I desire it'). But at first glance, when they are no longer self-deceived self-deceivers seem to discover what they did not know before (the young woman self-deceivingly who choose a scientific career not moved by her own interest but in search of her parents' approval, later discovers her true inclination for fine arts). See the path sketched by Hacker (1996)'s to dismiss SD as a counterexample to Wittgenstein's thesis.

SD' cases. Notice that if SD is intentional and X knows the true rational import of the available evidence, then SD is dissociative too. The intention to believe is the source of the belief that p, and the evidence the basis for holding that not-p. However, the reverse is not always true. If X believes that not-p and that p, it may be that he does not voluntarily hold the conflicting beliefs.[10] Now I know of no theorist who admits SD's dissociative character but denies its intentionality. The reason is that, once SD's dissociative character is acknowledged, the best explanatory rationale for X's holding that p is the interested intention to do so. Also, for the interpersonal model theorist, three necessary conditions have to be fulfilled for X to be self-deceived about that p:

(4) X believes that not-p.
(5) X believes that p.
(6) X believes that p because X intends to believe that p.

(4) and (5) define SD's dissociative nature and (6) its intentionality. In support of the requirement that (4) self-deceiving behavior is usually alleged, suspiciously eluding the evidence. Think of the woman systematically avoiding all confrontation with the data suggesting her husband's infidelity and continuously declaring to her friends her incapability to tolerate such a loss of trust. There seems to be a tension between the tendency to believe that p (verbally expressed) and the knowledge that nonetheless the evidence favors that not-p (not verbally expressed). Given the strength of the evidence against the verbally expressed belief and her non verbal expression of intentional avoidance, some conclude that in most cases, although X knows that not-p, she has managed somehow to manipulate her mind in order to believe that p. X has a strong motive to disbelieve that not-p, so that to escape the pain of the acknowledgement of the unpleasant truth, she devices a plan to conceal it. The argument favoring that (6) is this: if the self-deceived subject sees the evidence as menacing his mental equilibrium or his happiness,

[10] Some would say that there is no other way for acquiring beliefs, that they are not subject to our will. I shall address this point later.

then he has to know that the stubbornly declared belief that p is really false and succeed intentionally in not facing its persuasive power.[11] I shall call it 'the perverse blindness' argument, as it sees SD as a kind of voluntary blindness in face of the evidence.

Later on I will dwell on the intentionality question but now I would like to focus on SD's most famous feature: its seeming inconsistency or dissociative synchronic feature.[12] Logic prevents that conflicting beliefs are both true at the same time, but only the Davidsonian charity principle forbids us to ascribe them simultaneously to a rational being. According to this normative psychological principle only beliefs and desires maximizing rationality are to be ascribed to a rational subject. Attributing to someone the holding at t of the belief that p and that not-p counts *prima facie* against this principle, and is also an infraction of the more specific 'principle of the exclusivity of belief', according to which a proposition is believed at the expense of its rivals, so that to believe that p implies rejecting alternatives to p. As the charity principle advices us not to attribute to the subject at least serious breakings of the rationality requirements, like the holding of obviously conflicting beliefs, supporters of the interpersonal model for analysing SD have to work out a way to harmonize the charity principle with the attribution of SD to a subject.

To some, however, the attribution of inconsistency does not seem to be a troublesome question. After all, cases like the preface paradox allow us to diagnose an inconsistency feature common to most rational beings. To account for SD's inconsistency would then be no more awkward than to account for these cases. Let us look at this in more detail. Elster (1979, p. 178) is one of those inviting us to consider SD as one among these inconsistency cases. In the preface paradox the author avows that some of the statements made in the prefaced book are false. The

[11] This seems to be Davidson (1985)'s argument for (6).

[12] Synchronic dissociative SD involves that X simultaneously holds both the belief that p and that not-p (no matter whether one of them is unconscious or not). In diachronic dissociative SD cases X holds the conflicting beliefs at different times.

thing is that every reasonably modest person holds tacitly such a metabelief about his whole set of beliefs. Now the whole set of his beliefs would prove to be inconsistent, given that the truth of this metabelief is not compatible with his believing each of his beliefs to be true (some of them would have to be false). But if the whole set of my beliefs is inconsistent, then I believe everything, no matter how false or even absurd, given that from an inconsistent set of sentences every sentence is implied. The dissolution of the paradox in Elster comes from a distinction between generic and specific sentences, but the preface's author does not believe that the same specific sentence is both false and true: he does not simultaneously believe that p and not-p, but in the truth of each of the specific sentences of his doxastic web and in the generic sentence expressing the metabelief, unaware of the specific sentences which contradict it. This type of inconsistency, by no means regrettable, denotes no more irrationality than is to be foreseen in the framework of imperfect human rationality.

Now can we capitalize on the Elsterian strategy to eliminate the problematic inconsistency in SD cases? I do not think so. Unlike the preface case, in SD inconsistency is not between generic and specific sentences, but between two specific sentences (X sustains the truth and falsity of the same specific sentence 'p'). So, the inconsistency remains in its intolerable version.

Some authors who do not seem concerned by attributions of inconsistency go one step further in their disdainful attitude to the charity principle. Rorty (1972, pp. 393–396) sets out as requirements for SD that the subject not only believes that p and not-p, but that he recognizes, on the one side, that it is not rational to have incompatible beliefs and, on the other, that he believes that there is a strategy that reconciles his believing that p and his believing that not-p. The subject, then, would be doubly irrational, for he would have not only conflicting first-order beliefs, but conflicting second-order ones also, because of his believing in the possible compatibility of his incompatible beliefs. In my view there is no need to ascribe this hypertrophic inconsistency: the mother who deceives herself about her son's morality does not need to believe that there is a way to reconcile the son's real delinquency with his desired innocence. What she actually would like to reconcile would be her (false) belief and desire of her child's innocence with the world, which so stubbornly contradicts them.

2. *Strategies to Avoid Inconsistency*

This far, supporters of the interpersonal model are obliged to advance one of the following strategies, in order to avoid the attribution of inconsistency to the subject:

(a) the conflicting beliefs involved are partially or totally isolated. The weak version of this strategy takes as unconscious or unnoticed one of the beliefs (the unconscious or unnoticed belief strategy). Hard-line supporters of this option take the strong version, which postulates that each belief is modularly separated from the other in distinct compartments or subpersonal levels (whether conscious or not), so that both beliefs are inferentially encapsulated (the compartmentalization strategy).

Demos (1960) favors the weak strategy. Different versions of the strong one may be found in the ego-id-superego Freudian narrative, Pears (1982)'s divided mind theory and Fingarette (1969)'s splitting of the ego account. Pears advances that an accurate analysis of SD demands the existence of two rival centers of mental activity, defined by the magnetizing power of two occasionally conflicting desires: the desire for truth and the desire for pleasure or happiness (their similitude with the Freudian reality principle and pleasure principle should not pass unnoticed). Accordingly, the mother's unpleasant belief about her son's guilt would be magnetized by the activity center ruled by the desire for truth, but her SD would lead her to acknowledge exclusively the pleasant contrary belief, stocked in the rival mental nucleus. Unlike Pears (1982) and (1984, chap. 5), which is committed to the unconscious nature of the truth desiring center, Davidson (1982) presents a proposal demanding a porous compartmentalization of conscious mental centers. I shall leave the exposition of my reasons to distrust the exotism of mental ontologies like these until later.

(b) the propositional attitudes about the conflicting beliefs involved are different (the compatibilist strategy).

This strategy may be advocated by those defending some kind of 'epistemological separatism', on the basis of which the linguistic evidence admits the possibility of knowledge without belief;[13] too revisionist for me, but not for Haight (1980), who claims that the person self-deceived about that p believes that p, but nevertheless knows that not-p. The obvious advantage of this account is that, from the interpersonal model of analysis and without any commitment to compartments, the seeming inconsistency of SD disappears: when I believe that p and that not-p simultaneously, I exhibit a contradictory mental state, but not when I believe that p and know that not-p, provided that knowing that p does not imply believing it, as the epistemological separatist admits: an excessively high prize to pay.

(c) the conflicting beliefs are held at different times (the temporally restricted strategy).[14]

(d) the conflicting beliefs are partially held by the self-deceived (the partial beliefs strategy).[15]

Strategy (a) has been by far the commonest response to the problem posed by the compatibility of the conflicting beliefs thesis and the charity principle (what Mele (1997, p. 92) calls 'the static puzzle of SD'). It is clear that the problem does not arise for the interpersonal case, from which the SD's analysis is built up. The double instance of deception makes logically possible the consistency of the belief that p with the belief that not-p, because they are held by different subjects (the deceived and the deceiver respectively). It is psychologically explained that I am deceived, as long as the trust put in the deceiver prevents me from considering all the relevant evidence necessary not to be deceived. In the intrapersonal case, where only one instance is involved, if it is the same subject who holds

[13] See Luper-Foy (1992, pp. 234–235)'s presentation and comments on some of the defenders of this view.

[14] Sorensen (1985). I shall refer to this strategy when addressing Zamir's diary case.

[15] Gibbins defends this account in his comments to Mele (1997).

that p and not-p, who assesses and does not assess the relevant evidence, then the situation seems inconsistent.[16] So, every interpersonal guided analysis of SD finds one main source of the irrationality of SD in inconsistency. Interpersonal analyses are not obliged, however, to claim the subject's full awareness of his mental inconsistency. The lack of awareness points to an unconscious belief. The unconscious or preconscious inconsistency to which the compartmentalization theorists appeal prevents the flagrant violation of the charity principle. The self-deceiving mother will believe unconsciously that her child is a delinquent, but she will claim consciously and sincerely that he is not. The respectability of this strategy depends on the definition of an identity criterion for unconscious mental states. This is a tricky demand, because of the *sui generis* nature of the identification conditions of these mental states, not inferable by definition out of verbal behavior criteria. If X claims that he believes that p, we infer then that he believes that p and that he is conscious of his believing.[17] Notice moreover that unconscious beliefs do not satisfy the exclusivity principle of belief: the holding of the unconscious belief that p would not imply the impossibility of holding another unconscious belief that not-p. To what extent are they really beliefs, and not, say, proto-beliefs? Given that unconscious beliefs are verbally unexpressable, non-verbal behavior cues are searched.[18] Some supporters of the unconscious belief thesis will think that SD is somehow analogous to blind vi-

[16] My research aims are mainly ontological and it is not my purpose to deal with ethics here; nonetheless, let me just remark that an ethical paradox supervenes on the epistemological paradox described here. You may be accused of believing counterevidentially that p. However, as you believe that not-p, you are blameless. Think, e.g., of some self-deceived Germans living during the Second World War, who were aware of the suspicious disappearance of their Jewish neighbors (Elster (1979)'s example). As we shall see, none of these paradoxes arise from a non interpersonal analysis of SD.

[17] "A linguistically expressed belief is a conscious belief" notes Peacocke (1992, p. 154).

[18] Sackeim and Gur (1979) display and interpret some experiments designed to prove the existence of such unconscious beliefs in SD cases.

sion cases and the phi phenomenon with different color spots.[19] Nonetheless, as far as I can see, any experiment with SD cases can be accounted for without the attribution of an unconscious belief conflicting with the verbally expressed belief. Even when the verbal behavior strongly suggests that the subject believes that not-p, it is always open the possibility of interpreting his claiming that p is insincere. The verbal behavior does not exhibit enough articulation to determine an identification criterion of a belief state: to behave *as if* one thought that p is not to behave *expressing* the believing that p. In any case, though I shall not be concerned with elaborating an identification criterion for SD, but rather with investigating the ontological identity question, it seems clear (from a minimal verificationist point of view) that the reliability of the unconscious beliefs strategy is jeopardized by this identification indeterminacy.

In its strong version this strategy has to appeal to subpersonal modules, mental sections where the conflicting beliefs and other linked mental states remain inferentially isolated. A sketch of what 'inferentially isolated mental states' means here may be offered as follows. Affirmative sentences p1, p2, p3. . . pn expressing mental states are inferentially isolated provided that the subject possessing the corresponding mental states may be aware that p1, p2, p3,. . . pn, but has no awareness of some of their Boolean combinations (e.g., p1 and p2, p2->p3,. . .) The self-deceiving mother would be said to know her belief that her son is innocent (p1) and that she implicitly knows he is not (not-p1), but she is unaware that she believes both (p1 and not-p1).

The compartmentalization strategy is a theorical device of unmistakable ancestry: Plato uses it to tailor his tripartite soul theory (*Republics* 439e–440a). Grounded on the motto "Divide

[19] See Dennett (1991, p. 114 and ff.) and his alternative analyses of these phenomena. He defines two possible interpretations: the one given by the Stalinist reading and the corresponding to the Orwellian reading. The Stalinist reading of the SD case would be that the belief consistent with the evidence is unconscious and that the self-deceiver revises his perception of the facts (this is the weak version of strategy (a)). According to the Orwellian reading, the SD would be wholly conscious of the correct belief, but will modify eventually his recollection of it by working out an epistemologically perverted plan (strategy (c)).

and rule”, it is conceived to surmont the conflicting beliefs thesis by reproducing in the intrapersonal deception case the non inconsistent duality of the interpersonal case. Now, despite its face value for resolving other mental conundrums, the thing is whether the compartmentalization strategy is unavoidable for SD theorists, insofar as its use is not wholly uncontentious. Moreover, I think it shows some important flaws. On the one side, it seems to deny the true possibility of SD: it is not X who is really self-deceived, but subX1, who by believing that not-p and feeling unpleasant its truth, intends to carry out a project to deceive subX2, the true deceived entity, in order that he comes to believe that p. On the other, it postulates an enduring splitting of the mind in order to account for a usually circumstantial though persisting mental state. SD is actually a dispositional state, i.e., it is not episodic (as a disease it would be closer to tuberculosis than to epilepsy), but its duration in the non-pathological instances does not require the permanent presence of compartments, which seem *ad hoc* explanatory devices. More than that, compartmentalization does not solve the ‘dynamic puzzle of SD’:[20] how a subject intentionally adopts a new belief and manages to hide this very intention from himself. Homunculi like subX1 and subX2 simply restate the paradox by moving only one step back. Now homuncular modularity appears to be the only available answer to this puzzle from the interpersonal model of analysis: if there is an intentional project, the engineer homunculus has to possess all the relevant information, but then, as the whole person, he cannot manage to hide the project to himself.[21]

[20] Again this is Mele (1997, p. 92)’s terminology. See also Elster (1979, p. 77).

[21] Let me recall that my scruples are about homuncular modularity, not simply about modularity, a successful strategy to solve other problem cases in psychology and philosophy of mind. Some theoreticians —Searle among them— have argued against defenders of the strong IA thesis with the objection of the homuncular fallacy. A way out of it in that case may be transferred in profit of the SD case, but it is fatal to the interpersonal model. From a Dennettian gradualist point of view, the human machine’s intentionality is not due to the efficacy of intelligent mental homuncula but to mechanical non intentional subsystems gradually more complex and well coordinated. Analogously

As expected, the approach of SD I shall favor is by no means compartmentalized, nor based on an interpersonal model. Some theoreticians, like Mele, do not reject as invalid *tout court* the interpersonal model, but think there are in fact instances of strong self-deception, even though they are not prototypic;[22] some others, whom I include myself, do not acknowledge the existence but of 'weak SD' cases (neither dissociative nor intentional, but motivational: the acquired belief that p results from the causing unpurposeful desire that p). I shall try to show that there is no empirical need to recognize the existence of strong SD cases and that interpersonal model analyses are empirically and methodologically unsuitable.[23]

3. *The Inferential Blindness Theory of SD*

Let us now move on and display the account I shall favor. By way of introduction, some remarks on the logical and ontological nature of SD are necessary. A case of SD is a mental enduring trope, i.e., a mental abstract (to exist, it depends on the concrete mind or person suffering from it) particular (temporally located) persisting by being exactly the same during a time interval. It is a doxastic state, so it has a propositional object (there is no simple

the dynamic puzzle of SD cannot be solved by appealing to an unconscious project plotted by a subpersonal homunculus with personal capacities, but to non intentional subsystems accounting for the apparent (but, in this case, not real) intentional nature of SD. The motivational theory I shall defend rejects the real intentionality of SD by a similar line of reasoning.

[22] For Mele (1997) prototypic SD cases are neither intentional nor dissociative, but biased motivational states closely related to wishful thinking phenomena.

[23] Among the critics of the interpersonal model analysis admitting only weak SD cases are Siegler (1968), Szabados (1973) and Bach (1981). Notice that Siegler (1968) do not offer a motivational account of SD. Canfield and Gustavson (1962) put forward an approach rejecting the interpersonal model and the dissociative nature of SD but accepting its intentional nature. According to them, SD is a kind of self-command. It consists in making oneself believe or forget something against the available evidence (ibid., p. 33). It can be objected against their proposal that (i) it does not allow to distinguish between SD and simple mistake, and (ii) it does not specify how can it be that I do command myself to *believe* (nor merely to *act*) insofar as beliefs do not seem to be will-dependent.

or direct SD by analogy with simple seeing versus seeing that): to be self-deceived about p is to believe (in a certain irrational way) that p. SD is the end product of a process. No instantaneous SD is possible. On the one hand, at least a temporal interval is needed in order that the desire cause the counterevidential belief. On the other, SD is a recalcitrant mental state, bringing about a belief retention in face of the contrary evidence.[24] Insofar as it is a state, it is dispositional, not occurrent: its cognitive nucleus is not a thought but a belief. I take, as usual, that thoughts are episodic and known to their possessor, whereas beliefs are dispositional and most of them remain unexpressed and even unknown by their own holders (think, e.g., of your belief that whisky is not solid or that non-Carrollian caterpillars do not smoke).

Although SD is always self-centered (see below (vii)), it may or may not be egocentric. Egocentric SD is always about states of affairs one of whose essential *relata* is the self-deceived (his mental states —the priest's unconfessable love for a non divine entity— or his non mental states —the drug addict's unacknowledged physiological dependence from drugs). Non-egocentric SD is about other's mental or non-mental states, where the self-deceived subject is only an accidental constituent. As a doxastic state, SD may have as objects emotions, desires, beliefs, abstract or concrete,[25] and be of first or n-adic order (I may be self-deceived about my own SD).

On the approach I defend, the 'inferential blindness theory of SD', given a rational, inferentially competent and autonomous being (i.e., with normal inferential faculties and not under the effects of drugs, hypnosis and so on), X, and a proposition, p,

[24] SD may be, however, punctuated by episodic moments in which the subject entertains the proposition that not-p without giving assent to it. Some SD cases may have an intermitent existence. This is what happens when the self-deceived mind shows episodes of lucidity during which not-p is not only entertained, but actually thought of as true.

[25] It is somehow surprising that Mele (1997, p. 95) indicates that he is only concerned in analysing cases of *belief*'s acquisition/retention, as no other possibility remains. The belief's content may change from case to case (and be an emotion, a desire. . .), but SD is nevertheless a doxastic (not an emotional, conative. . .) state.

if X self-deceives himself about that p from time t to t′ (t<t′), the following (generally) necessary and sufficient conditions are to be fulfilled:

(i) the rational assessment of the available evidence should lead X to conclude that not-p, and X knows or suspects the existence of this evidence.
(ii) X desires that p.
(iii) X does not infer that not-p from the evidence, because of the interference from his desire that p (inferential blindness).
(iv) X believes that p.
(v) X believes that p because X desires that p (this desire is the proximal cause of the belief)
(vi) X does not know that (v).
(vii) X′s belief about that p is a self-centered belief.

This account does not take deception but wishful thinking as the model for analysing SD (as (v) states). This is a deflationary account, so that it is not committed either to theoretically expensive architectures as Freudian creatures or subpersonal compartments and unconscious intentions. It denies SD's alleged intentional and dissociative nature and accepts only weak motivational SD cases. SD is basically understood as a kind of counterevidential motivated belief, whose irrationality is not rooted in its inconsistency but in its ill-grounding (a desire is not the right kind of reason for holding a belief).

Let me have a look on the requirements (i)-(vii). The qualification in (i) demanding that X knows or at least suspects the existence of counterevidential data is required to distinguish SD instances from instances of unmotivated mistake. Actually (i) states (in the first part of the conjunction) the counterevidential nature of SD and (in the second part of (i), and jointly with (v)) draws the difference between SD and mistake by ignorance, i.e., mistake due to unattentiveness or surface outlook. In the former but not the latter case the held belief is due to the desire's causal role and the subject knows the evidence, in spite of the fact that he does not recognize its true rational import against the sustained belief. The wife is acquainted with the evidence that her husband comes home later than usual, that he is often in a bad mood [. . .] but symptoms of her SD show in her attributing the etiology of his change of behavior to other less worrying causes. What happens is that the desire of her husband's loyalty produces in the woman a kind of inferential blindness which prevents her to extract from the evidential data the more dramatic consequences. Recall that friends of the interpersonal model pointed out to the perverse nature of SD's blindness. Now I do not think this perversion —as intentional and biased— is rightly ascribable to the self-deceived. Their argument ran as follows: as the self-deceived avoids the acquaintance with the evidence, he has to be aware of its menacing import, then he has inferred the unpleasant proposition but perversely conceals it from himself. Is it convincing? Obviously, if the evidence is eluded *because* it is seen as menacing, then it has to be recognized *as such*. The thing is, however, whether the evidence is eluded or simply unknown, or whether, though known by the subject, he does not see the rationally inferrible consequences. I bend rather in favor of the latter option. On my account, the desire's causal strength produces a kind of focussed inferential blindness in the subject's mind, so that, even when openly facing the evidence, he is unable to conclude what another subject (and even he himself) in a dispassionate mood could do.[26] So far then, the self-deceived addresses directly to the belief he desires to be true; he does not form it first and later rejects it because of his contrary desire as a direct outcome of an epistemologically perverse planning. The young philosopher submitting a paper to the scientific board of a journal of repute is wishing it will *be* published, not *believe* it will be. The strong desire leads him to undervalue the paper's flaws and overvalue its virtues: SD is a kind of selective exposure

[26] For a desire to undermine the subject's local inferential capabilities it may appear necessary that desire was bounded to emotion. Emotions are typical suspendants of the rational faculties, sometimes as reason disturbers, sometimes as reason enhancers. Could it be that a dispassionate uncolored desire produced a self-deceptive focussed inferential blindness? Actually, I do not know whether there are uncolored desires, as there are uncolored beliefs and judgements. Fortunately, my main point does not hinge on the answer to these questions.

to information, carrying an inferential dysfunction temporally circumscribed to the relevant situation.[27]

SD's focussed inferential blindness is double-sided. Firstly, it is extrospective blindness, for the self-deceived subject is unable to make the inference strongly suggested by the known evidence. Secondly, it is introspective blindness. The subject is ignorant of the causal link connecting his desire that p with his belief that p. So the link is temporally inaccessible to the self-deceived consciousness, "inaccessible to the machinery of reasoning and action control". This ignorance counts as a paradigmatic counterexample to the Cartesian thesis of the mind's transparency and introspection's incorrigibility.[28] No privileged authority can be attached to ascriptions of these first-person mental states. Indeed, phenomenological evidence, available solely to the subject, is no help to attain insight in SD cases. On the contrary, extern witnesses use to diagnose earlier and better. The blindness involved, however, is by no means voluntary, i.e., the result of an alleged decision to believe counterevidentially. I find no explanatory need to attempt like this against the charity principle. Certainly someone should characterize in detail the mechanisms associated with what I term 'inferential blindness'. But I am unsure whether this is a task for a philosopher or a psychologist.

Maybe an analogy from the philosophy of science can be useful here. In certain sense, the desire that p in the self-deceived performs the function of the Lakatosian protecting belt saving the scientific theory from the revision *prima facie* demanded by the counterexemplary evidence. The self-deceptive mental behavior would then be like the XIX-scientist supporting the phlogiston theory. To explain the metal combustion and in face

[27] Remember that no SD state about that p is ascribable to someone medically disabled in general to infer that not-p from the data at his disposal.

[28] During much of her life a woman sincerely avows to choose the medical career out of her own preferences. For fear of losing her parents' esteem, she neglects her artistic gifts, which they undervalue. Later she discovers the true reason for her choice. Overcoming her SD will show her that she did not have an infallible knowledge of her belief state (because it could be *shown* that she was wrong), nor incorrigible either (it *was* wrong). Shoemaker (1994)'s use of the terminology.

of the evidence that it gained weight after combustion, he did not conclude that combustion cannot consist in a phlogiston emission but suggests that phlogiston has negative weight. Self-deceptive beliefs are conservative *ad hoc* hypothesis, grounded in the unjustifiable modifying of auxiliary conditions. Unlike the *ad hoc* hypothesis about Neptune's existence, SD evidence is not compatible with the original belief, but it persuades us to substitute it for its contrary as an anomaly.

The self-deceptive belief, then, is acquired as the effect of a proximal or direct cause which is the desire, not of believing that p, but that p was the case (condition (v)). This desire, in turn, can be the causal effect of an emotion, which in turn would be the distal or indirect cause of the self-deceptive state. In Wilde's case, love grounds the desire to be requited by his beloved and this desire may make him believe this is so in the teeth of the contrary evidence reported by himself.

As said before, something distinctive of SD is that the belief's acquisition and persistence is stubbornly opposed to the available evidence,[29] but not that the acquired/persisting belief be false. Notice that among (i)–(vii) there is no necessary condition stating that 'p' has to be false. I regard such a requirement as an unusable reminder coming from the false analogy deception-SD.[30] To be self-deceived about that p, like to wishfully think that p, is a matter of what can or cannot be rationally inferred out of the evidence, independently of the truth or falsity of 'p'. SD is a matter of epistemic reliability, not of truth value. After

[29] In SD an intensification of the doxastic natural inercy is produced. Peirce told us that doubt is an uncomfortable and unsatisfying state of mind, that we seek to release from it and strive to recover the comfortable state of belief. However, our tendency in search of doxastic stability is sometimes broken by the intervention of our tendency to acquire true beliefs. The self-deceived exhibits an epistemic inertial state that neutralizes the potential strength of this latter tendency.

[30] Mele (1987) and (1997) demands, nevertheless, that it be so. Mele (1997, p. 95)'s rationale for requiring in SD cases the falsity of 'p' is this: "a person is, by definition, *deceived in* believing that p only is p is *false*; the same is true of being *self-deceived in* believing that p". It is quite surprising that he accepts here the analogy deception-SD when he refused it formerly in his general motivational account of SD.

all, the son of the self-deceiving mother could be innocent, despite the evidence to the contrary. The logical difference between the verification conditions of a sentence and its truth conditions counts as an argument against the inference from the first part of (i) to the falsity of 'p'. A counterfactual remark could help us to understand what I am trying to explain: if no extra-evidential source supported the mother's belief except her desire of the child's innocence, the occasional truth of her belief would not afford us to deny that she was nonetheless self-deceived about her son's innocence.

As a rationale for (vii) I shall claim that no one self-deceives oneself about a proposition which does not occupy a place in his evaluative-emotional space. If I am not personally concerned by its truth or falsity, then p is not a candidate object for a possible SD of mine. In this personal concern (whether it is known or not by the subject) consists what I call the 'self-centered' nature of the self-deceptive belief. I disagree, however, with Rorty (1972) and Taylor (1985, pp. 120–123) about the role they assign to personal identity in SD cases. The variety of ordinary and trivial SD occasions dissuades me from the temptation to magnify: not every SD instance is concerned with changing or preserving one's personal identity or integrity. Certainly in the young poet's self-deceptive case his self-esteem is at hand. But cases such as the self-deceiving mother need not be cases where the mother is feeling responsible for her wrong bringing up, but they may be simply motivated by a desire to the best for her son.

4. *A-Intentionality and B-Intentionality*

Recall that among the negative reasons to adopt a motivational model for analysing SD there are some related to the solving of the dynamic puzzle, i.e., related to SD's alleged intentionality. Sartre (1943, part 1, chap. II) seems to have been the first to put it clearly: if X comes to believe that p by desiring to believe, how is it that he succeeds in executing the project? As no answer is compelling, Sartre rejects SD's possibility as pragmatically impossible to work out: I must know the truth very exactly in order to conceal it more carefully from myself. My response to the dynamic puzzle is its dissolution: there is no logical or pragmatic difficulty in achieving the project, because there is no project at all.

Now, in a certain sense SD is uncontroversially an intentional state, as it is a doxastic state of mind and every belief is intentional (B-intentional) in the Brentanian sense of representing or pointing to an object. SD's B-intentional nature explains the familiar phenomenon of referential opacity diagnosed in the sentences involved in the reporting of SD states. 'Moosbrugger's mother believes that the prostitute's killer is a rogue', 'Moosbrugger's mother does not believe that Moosbrugger is a rogue' and 'Moosbrugger is the prostitute's killer' would all be true, if Moosbrugger's mother were self-deceived about her son's innocence.

But, although SD is B-intentional, I do not think it is A-intentional, i.e., intentional or intentioned as an action is, *pace* Davidson (1985), who takes that SD is the result of achieving a plan previously worked out for convincing oneself that p, though knowing that not-p. Davidson establishes an analogy between belief and the decision to believe, on the one side, and acting and the decision to act. This view has a Cartesian flavor, in agreement with the thesis that beliefs are subject to the will, and opposed to the Humean view which understands them as passive and not will-dependent. As explained later, I will not take Hume's nor Descartes' part, but defend SD's non A-intentionality for explanatory reasons. Against Davidson and his view of SD as perverse blindness (see condition (6)), (v) and (vi) try to capture the idea of this involuntary aiming.

It is worthwhile to remark, by the way, that an argument has been put forward to deny not only SD's empirical intentionality, but also its conceptual possibility. The argument is based on the Humean source to which I have just been referring. The belief involuntariness thesis establishes that no belief can be, by definition, formed as an outcome of a voluntary plan. It is advanced by Williams (1973, p. 148) and Williams (1978, chap. 6). I can decide to act against my best judgement (akrasia), but this possibility does not follow for belief. Belief is not a matter of choice. *A fortiori* no self-deceptive state can be achieved by planning it, as it is logically, nor merely pragmatically, impossible to do it. Certainly most of our beliefs are empirically formed by unintentional means, but I would not dare to say that *all of them* are. Some exceptions, like successful Pascalian projects, count as counterexamples for Williams' thesis. In tune with Elster (1979)'s remarks on by-product states, I shall defend that, even if it is possible to believe intentionally that p, it cannot be done using direct strategies, but by means of indirect ones. Beliefs may be, then, either involuntary or indirectly voluntary. Let me develop this point more carefully. Think of Oscar, who desires to forget his love for an unattainable man. He may conceive of a Pascalian project to succeed in effacing his memories of his. He goes away from the town where he met him, destroy the material things reminiscent of their meeting, arrange some feminine rendezvous and eventually marry a woman and have children. Can he succeed in his forgetting project? Some manage it, but one condition is required: the person involved must formerly forget the self-deceptive plan. Then the resulting belief (in that case, that he no longer loves him) will be indirectly acquired, as the project's by-product. Elster reminds us that some actions, like sleeping or spontaneous behaving, are essentially indirectly achieved, so that it is impossible to achieve them as a result of direct strategies. Most of us know that strategies of sleep induction work only as long as we are unaware of the planned operation, because the project's awareness would block its performance. By analogy, during the process of the belief's acquisition the knowledge of the project would causally interfere in its achievement. So far SD could be *prima facie* A-intentionally attained as an indirect by-product, provided that the project is concealed for the subject when SD is taking place. It may be a hard mental job, but not a conceptually impossible one. Notice that a characterization of SD as A-intentional and synchronically dissociative would demand that the concealed project was unconscious. However, if SD is A-intentional but diachronically dissociative the concealment is only necessary during achievement, not during planning. Of course, the planning exhibits the irrational character typically attributed to self-deceptive states: the subject knows (or truly believes) that not-p, but he plans to believe contrarily.

This seems to go against the belief's exclusivity principle and the belief's truth aiming nature (to believe something is to believe-it-true). So how can it be that I believe something to be true while I believe it false? Actually the subject does not plan to believe that p and know that not-p, but to believe that p and forget that he previously believed the contrary.

From the nature of the belief, then, no conceptual objection seems to preclude A-intentional SD. So, why does my account not accept strong SD cases? Let us see a seeming instance of A-intentional diachronically dissociative SD: Zamir's diary case.[31] Imagine that Zamir projects at t to deceive his future self at t′ (t<t′) by writing a wrong description of an upsetting meeting with a close relative of his. As time passes he forgets having introduced the wrong entry and acquires the false belief that the meeting happened that way, as planned by his past self. Is this not a true A-intentional instance of SD?[32] Not at all, insofar as (i)–(vii) be SD's requirements. At least three of SD's necessary conditions, namely, (i), (ii) and (v), are not satisfied by Zamir's diary case.[33] First of all, when future Zamir reads the diary, his belief does not go against the available evidence, so that his mental state is not epistemically irrational. And secondly, his belief is not formed by the pressure of a desire to be true, no desire drives him to misinterpret the available data favoring the contrary belief. Finally, the proximal cause of his belief is not a desire. It is true that he has somehow induced himself to mistake, but he is not really self-deceived. Oscar's project to forget his beloved is not a true instance of SD either: his final belief results indirectly from his desire to forget and he exhibits no evidential blindness.

[31] Based on Mele (1997, p. 99)'s example. See also Davidson (1985, p. 145).

[32] Recall the temporal restriction thesis (strategy (c) for solving the problem posed by the conflicting beliefs thesis). If all SD is diachronically dissociative, like Zamir's diary case, then no inconsistency is ascribable to self-deceivers, because they hold the conflicting beliefs at different times. Notice, however, that this strategy denies the basic deception-SD correspondence, as diachronic instances do not fulfill the requirement (3) of deception.

[33] Mele (1983) presents the case of Guido, who in the end comes to believe in God by ruling a Pascalian plan, and qualifies it as self-deceptive.

So far, my account only accepts weak SD cases. The seemingly strong SD cases are explainable in terms of weak SD or are instances of induction to mistake. The current variety of self-deceptive instances can be accounted for with no appeal to intentionality and dissociation, but solely in terms of the causality of a desire and a consequently involuntary blindness to be aware of the true evidential import of the available data.

Now SD, though non A-intentional, is a motivational state of mind. Condition (v) states that the self-deceptive state is effected by the acting of the desire that p, which produces the belief that p. The mere existence of the desire is not enough to produce a self-deceptive state, nor acknowledging the existence of the desire suffices to overcome SD. The way out from SD is to acknowledge that the proximal and unique direct cause (a cause, not a reason) of the belief was the desire. This should be enough for a rational subject to reject his mental state as epistemologically deficient.

The idea of the causal link allows us to evade a possible objection to Mele (1997) and Szabados (1973), who hold that it is the mere existence of the desire that motivates the acquisition of the self-deceptive belief. To see the benefits of the causal link qualification, think of two subjects, X and Y, who share the same desire type (with the same emotional color) and belief type that p, but such that only X is able to revise his belief as he is aware of the relevant evidence, whereas Y suffers from focussed inferential blindness and is self-deceived. The situation described is clearly a possible one. But then the mere existence of the desire cannot be the SD's causal trigger, for it is present in both X and Y, and only Y is self-deceived. Supporters of the interpersonal model could use this to argue in favor of SD's A-intentionality: it is the intention to believe what is desired which is present in Y′s mind but not in X′s. On my view no A-intentionality is mandatory to solve the problem posed by that situation. Certainly the desire's presence alone does not determine the self-deceptive belief's acquisition. Most SD theorists[34] use to affirm that the self-deceived would acknowledge his state

[34] Among them Davidson (1985), Pears (1982) and Mele (1997).

as wronged in the desire's absence. But I take that neither the acknowledgement of the desire's existence is a necessary nor sufficient condition for the persistence of the subject's SD. It does not suffice for SD: although the self-deceiving mother may desire that her son was innocent, she may be clear-sighted enough to see the truth and that her desire should not interfere with her rational assessment of the evidence. Nor is it necessary: even if the mother knows of her desire, she can nonetheless be self-deceived, insofar as she does not realize that her desire is the only cause of her believing in her son's innocence. In fact, the difference justifying the distinct states of X and Y depends on the different functional role played by Y′s desire in the web of his mental states, which determines a self-deceptive effecting in Y, but not in X. As in any other causal relation, no effect is obtained in absence of the appropriate initial conditions.

5. *Concluding Remarks*

By way of conclusion, some notes to sum up and clarify the main tenets of my approach. SD:

1. is a motivated state of focussed inferential blindness. Unlike the subject merely mistaken by ignorance, the self-deceiver knows the evidence against his belief, his blindness on its true import is due to the unintentional causal efficacy of a desire and his SD has as its object a proposition occupying some place in the subject's emotional-evaluative space. SD, then, is not a case of voluntary, perverse blindness, but of involuntary blindness: it is closer to mistake states than to deceptive states. Although SD is not A-intentional, it is not free of some kind of responsibility. Because of its cognitive content (unlike a sensation or a physiological disturbance), SD is assessed as epistemically irrational or not, appropriate or not. As with beliefs in general, which are usually not directly chosen, I am somehow responsible for the correct entertaining of some of my metabeliefs, said to maximize the epistemic warrants of my cognitive states (a topic for the so called 'ethics of belief').

2. even if its etiology is conative as in the case of wishful thinking, it is distinguished because in the latter case the evi-

dence is neutral and may be unknown by the subject, whereas in SD the evidence is by all rational lights against. Unlike simple mistakes, both have in common their being mental states (i) lacking well-groundedness, for in both a desire is the basis of the held belief, and (ii) are definable in terms of epistemic reliability, not of truth value.[35] Against the assimilation with deception, which cannot be achieved about a true proposition, I can have true wishful thoughts and be self-deceived about a true one. Now the wishful thinker does not need to know the relevant evidence for his wishfully acquired belief. Unlike SD cases, it is not present the conservative trend reluctant to the rational evaluation of the overwhelming evidence to the contrary.[36]

3. is not a kind of deception (the self-deceiver is not lying to himself), given that the self-deceiver about p does not know or believe that not-p (SD is not dissociative). The main reason why SD is seen as a paradoxical state of mind is the wrong assimilation of SD with deception demanded by the interpersonal model ('X self-deceives himself about that p′ is read as 'X deceives X about that p'). The inconsistency attribution is explanatorily superfluous and controversial regarding the charity principle: in order to be self-deceived about that p, it is not necessary to believe that not-p, but to desire it and that leave the desire cause the corresponding belief. SD's irrationality is epistemic, not logical irrationality: the self-deceiver is not holding synchronic nor diachronic conflicting beliefs.

The self-deceptive state may be accounted for by resorting to the unacknowledgement of the illegitimacy of the belief's causal

[35] Against Szabados (1973). For him wishful thinking is always about a false proposition. But I think that the irrational character of wishful thinking is due to the inappropriate unbinding of the belief's acquisition process from the available evidence. The naive romantic may be fairly confident about that his love is requited, solely by the force of his desiring it. The epistemological etiologies of the fanatic and the person experiencing *Schadenfreude*, on the one hand, and *mutatis mutandis* those of the congenital pessimist and the obsessively jealous ('counterwishful' thinkers according to Elster (1999, I.6)) may be explained similarly.

[36] This goes against Mele (1997), who thinks that wishful thinking is a species of SD.

source. If, consequently, SD fails to have a dissociative synchronic character, then mental compartments, *sine necessitate* pathologizers, are not mandatory. The compartmentalization strategy is not only guilty of the homuncular fallacy, but it goes against the charity principle, so that it is rejectable for methodological reasons.

We cannot usually decide to believe or disbelieve at will, but I know of no principle argument reasonably preventing that they are acquired as by-products of indirect strategies. SD, however, requires a direct grounding in a sustaining desire, which is not present in cases of alleged strong SD cases.

BIBLIOGRAFÍA

Audi, R., 1985, *Self-Deception and Rationality*, in M. Martin (ed.), *Self-Deception and Self-Understanding*, University Press of Kansas.

Bach, K., 1981, "An Analysis of Self-Deception", *Philosophy and Phenomenological Research*, 49, pp. 351–370.

Canfield, J.V. and D.F. Gustavson, 1962, "Self-Deception", *Analysis*, 23, pp. 32–36.

Davidson, D., 1980, *Essays on Actions and Events*, Clarendon Press, Oxford.

——, 1982, "Paradoxes of Irrationality", in Wollheim and Hopkins (eds.), pp. 289–305.

——, 1985, "Deception and Division", in Lepore, E. and B. McLaughlin (eds.), *Actions and Events*, Blackwell Publ., Oxford, pp. 138–148

Demos, R., 1960, "Lying to Oneself", *Journal of Philosophy*, 57, 18, pp. 588–595.

Dennett, D., 1991, *Consciousness Explained*, Brown and Co., Little.

Elster, J., 1979, *Ulysses and the Sirens*, Cambridge University Press, Cambridge.

——, 1999, *Alchemies of the Mind*, Cambridge University Press, Cambridge.

Festinger, 1957, *A Theory of Cognitive Dissonance*, University Press, Stanford.

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

——, 1982, "Self-Deception and the 'Splitting of the Ego' ", Wollheim, pp. 212–227).

Gardiner, P., 1969–1970, "Error, Faith and Self-Deception", *Proceedings of the Aristotelian Society*, LXX, pp. 221–243.

Guttenplan, S., 1994, "Self-Deception", in Guttenplan, S. (ed.), *A Companion to the Philosophy of Mind*, Blackwell, Oxford, pp. 558–560.

Hacker, P.M.S., 1996, *Wittgenstein Mind and Will*, vol. 4 of his *An Analytical Commentary on the Philosophical Investigations*, Blackwell, Oxford.

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

Kenny, J., 1963, *Action, Emotion and Will*, Routledge and Kegan Paul, London.

Luper-Foy, S., 1992, "Knowledge and Belief", Dancy, J. and E. Sosa (eds.), *A Companion to Epistemology*, Blackwell, Oxford, pp. 45–48.

Mele, A.R., 1983, "Self-Deception", *The Philosophical Quarterly*, 33, 133, pp. 365–377.

——, 1987, *Irrationality. An Essay on Akrasia, Self-Deception and Self-Control*, Oxford University Press, Oxford.

——, 1997, "Real Self-Deception", *Comments and Replies*, *Behavioral and Brain Sciences*, 20, pp. 91–136.

Peacocke, C., 1992, "The Concept of Belief: Self-Knowledge and Referential Coherence", *A Study of Concepts*, The MIT Press, Cambridge MA, pp. 147–176.

Pears, D., 1975, "Paradoxes of Self-Deception", *Questions in the Philosophy of Mind*, Duckworth, London, pp. 80–96.

——, 1982, "Motivated Irrationality, Freudian Theory and Cognitive Dissonance", Wollheim, pp. 264–288.

——, 1984, *Motivated Irrationality*, Oxford University Press, Oxford.

——, 1991, "Self-Deceptive Belief-Formation", *Synthese*, 89, pp. 393–405.

Price, H.H., 1965, *Belief*, George Allen and Unwin, London.

Rorty, A.O., 1972, "Belief and Self-Deception", *Inquiry*, 15, pp. 387–410.

Sackeim, H.A. and R.C. Gur, 1979, "Self-Deception: A Concept in Search of a Phenomenon", *Journal of Personality and Social Psychology*, 37, 2, pp. 147–169.

Sartre, J.-P., 1943, "La mauvaise foi", *L'être et le néant*, part 1, chap. II, Gallimard, Paris.

Schoemaker, S., 1994, "Introspection", in Guttenplan (ed.), pp. 395–400.

Siegler, F.A., 1968, “Demos on Lying to Oneself”, *Journal of Philosophy*, 59, pp. 469–475.

Sorensen, R.A., 1985, “Self-Deception and Scattered Events”, *Mind*, 94, pp. 64–69.

Szabados, B., 1973, “Wishful Thinking and Self-Deception”, *Analysis*, pp. 201–205.

Talbott, W.J., 1995, “Intentional Self-Deception in a Single Coherent Self”, *Philosophy and Phenomenological Research*, 55, 1, pp. 27–74.

Taylor, G., 1985, *Pride, Shame and Guilt. Emotions of Self-Assessment*, Clarendon Press, Oxford.

Williams, B., 1973, “Deciding to Believe”, *Problems of the Self*, Cambridge University Press, Cambridge, pp. 136–151.

——, 1978, *Descartes. The Project of Pure Inquiry*, chap. 6, Penguin Books, London.

Wollheim, R. and J. Hopkins (eds.), 1982, *Philosophical Essays on Freud*, Cambridge University Press, Cambridge.

Recibido: 6 de febrero de 2001
Aceptado: 2 de mayo de 2001