# Self-Deception: Intentional Plan or Mental Event?

*Anna Elisabetta Galeotti*[*]
elisabetta.galeotti@lett.unipmn.it

ABSTRACT

The focus of this paper is the discussion between supporters of the intentional account of SD and supporters of the causal account. Between these two options the author argues that SD is the unintentional outcome of intentional steps taken by the agent. More precisely, she argues that SD is a complex mixture of things that we do and that happen to us; the outcome is however unintended by the subject, though it fulfils some of his practical, though short-term, goals. In her account, SD is produced after a fashion similar to those beneficial social phenomena which serve some collective purpose, are the product of human action, but not of human design, such as money, language and many social conventions; and similarly SD can be accounted by invisible hand explanation. The paper will critically analyze both the intentional and the causal accounts, and then present the invisible hand explanation which avoids the most puzzling aspect of the intentional view, while keeping the distinctiveness of SD in the realm of motivated irrationality. A brief discussion of the issue of responsibility for SD will conclude the paper.

## Introduction

I hold that *self-deception* (SD) is believing that P against the available evidence and under the influence of the desire that P be the case. It is a form of motivated irrationality, displayed by usually rational subjects, capable to form and hold beliefs appropriately. In the discussion on SD developed in philosophy after Sartre's theorizing of *mauve fois* (1956) and more recently in various branches of psychology, the field has been contended between:

* Dipartimento di Studi Umanistici, Università del Piemonte Orientale, Vercelli, Italy.

a) skeptics and non-skeptics of SD as a genuine phenomenon;
b) supporters of an apparent paradoxical view and of a non-paradoxical view of SD;
c) intentionalists and non-intentionalists;
d) those who see SD as a culpable failure in cognitive capacities and those who consider SD as a vital response to difficult realities beyond the agents' control.

In this paper, I take SD as a genuine, puzzling but non paradoxical, phenomenon and I shall specifically focus my analysis on the intentional *vs.* causal account of SD. In this respect I shall defend the view that SD is the unintended outcome of intentional steps taken by the agent. I shall contend that SD is brought about indirectly by motivated mental acts elsewhere oriented. If it is the by-product of mental activity otherwise directed, then the subject's responsibility is likewise indirect: since SD is not simply a happening, but also a doing of the subject, the agent is not free from responsibility, but because SD is an indirect product, the responsibility concerns the failure to avoid being prey of SD.

If SD is considered a genuine phenomenon, that is, is not discarded as mere pretense and deception of others (Haight, 1980, 1985; Kipp, 1980, 1985; Gergen, 1985)[1]; nor as the normal outcome of cold biases (Gilovich, 1991; Piattelli & Palmarini, 1994; Friedrick 1994) or of brain modules lacking a unitary center (Kurzban, 2010), then the problem of whether it is something that we do or that happens to us is crucial. For if SD is a causal product of motivated biasing, then it is certainly non-paradoxical, nor especially puzzling (Mele, 1987, 1997, 2002) but in this case SD is conflated with various kinds of motivated irrationality, such as wishful thinking, illusions, faith, and also not well marked of from unmotivated irrationality such as delusion.[2] If, by contrast, it is viewed as intentional, then SD seems stuck in the "dynamic paradox" in so far as it seems logically impossible to bring about a false belief intentionally and cunningly in the teeth of evidence; the intentionalist has moreover to explain away the risk of the "static paradox" of the subject holding P and non-P, maybe

[1] There is another view of SD as pretense which still understands SD as a genuine phenomenon (Audi, 1982, 1988; Rey, 1985; Funkhauser, 2005; Szabò & Gendler, 2007).
[2] The risk of the conflation between SD and delusion is somehow acknowledged by Mele himself (2009, pp. 139–158).

via the problematic mind-partition.[3] Short of that, complex explanations, involving subconscious, mental tropisms, half-beliefs, and so on are then needed in order to account SD as an intentional, but non-paradoxical project.

As a way out, I shall argue that SD is a complex mixture of things that we do and that happen to us; the outcome is however unintended by the subject, though it fulfils some of his practical, though short-term, goals. I suggest that SD is produced after a fashion similar to those beneficial social phenomena which serve some collective purpose, are the product of human action, but not of human design, such as money, language and many social conventions which have been the focal issue for many economists and social scientists, starting with Adam Smith, and proceeding with Carl Menger and Friedrich Hayek. For this kind of phenomena, functionalist explanations have attempted to match the social purpose with a teleological scheme of explanation, where the purpose was either moved backward as a cause or ascribed to a presumed collective agent. Either way, the fallacy of such explanations have long been established,[4] and more satisfactory models, such as *invisible hand explanations*, have been proposed, showing that the beneficial effect is the unintended outcome of many individual actions elsewhere oriented and motivated, plus a processing filtering mechanism.[5] I see a clear analogy between phenomena produced by the invisible hand mechanism and SD: in SD, as well as in phenomena like money and market, there is a *purpose* which is served by the deceptive belief; and, if there is a purpose, it is only too easy to presume *a plan* designed to fulfill it, and *an agent* conceiving the plan and carrying it out. But, as in the case of beneficial social phenomena, the seemingly purposive outcome does not need to presuppose a teleological model to be made sense off.

In section 1 I will present the intentional account, pointing out its appeals and its drawbacks; in section 2, I will discuss the causal account which looks promising and apparently provides a response to the weakness of the rival view, but which exhibits different kinds of difficulty. In section 3, I shall argue that my invisible hand account avoids the most puzzling aspects of the intentional view, while keeping SD distinctiveness in the realm of motivated irrationality which is lost in the purely causal account. I shall conclude with a brief

[3] That there are two kinds of paradoxes involved in traditional views of SD, the dynamic and the static is clarified by Alfred Mele (1997, pp. 91–102).

[4] For a critique of functional explanation see Elster 1983.

[5] For a discussion of invisible hand explanation see Nozick (1974, 1977)

discussion of the problem of the responsibility for SD, as it emerges from the invisible hand view.

Before starting, I would like to preempt a potential objection. It may seem that my invisible hand explanation implying a beneficial outcome for the agent's (short-term) interests, only fits the so-called straight cases of SD, while it cannot make sense of "twisted cases" (Mele 1999; Lazar 1997, 1999). In twisted cases, SD purpose is not apparent, since the agent ends up irrationally believing what he does not desire to be true, hence the deceptive belief seems to run contrary to the agent's, even short-term, interest.[6] I think that invisible hand explanation could account also twisted cases, though I cannot pursue this point here. In any case, twisted cases do not constitute an obstacle to my view given that a unitary account has not yet provided a satisfactory explanation for either. Causal accounts of SD, most notably by Alfred Mele and Ariela Lazar, have actually stated that both types of SD are explained by their theory, and this seems to be an appealing feature which intentional accounts allegedly lack. But Dana Nelkin (2002) has shown that the unity comes with a price; Mele's view implies that the motivation triggering the causal biasing of data, ending up in the false belief, is content-unrestricted, so that the operating desire has actually no match in the deceptive belief. Hence twisted cases are explained by the same unitary model, but it is unclear that they are indeed SD cases. Nelkin's solution, by substituting the desire that P with the operating desire to believe that P (or in twisted case non-P), is far from being satisfactory, because then she has to explain why S, being usually rational, and having the desire that P, has nonetheless the desire of believing non-P. Supposing twisted cases are SD cases indeed, I think that a supplementary unraveling into *which* desires and *under what circumstances* can set off SD process is required for a possibly unitary account to be provided.

## 1. The intentional view

The intentional account of SD appeals to the intuition that the self-deceived subject (SDS) seems to display intellectual dishonesty in her conviction that P is the case despite one's contrary evidence. "Dishonesty" appears to be an intentional doing for matching her beliefs with her desires, instead of being rationally responsive to evidence. In turn, this leads to conceive SD as lying to

[6] The example made by Mele (1999) refers to the jealous husband who convinces himself, despite the evidence, that his wife is unfaithful, while desperately desiring her fidelity.

oneself, and to pave the way with paradoxes, namely the "dynamic paradox" of bringing oneself to believe that P, knowing non-P, and the "doxastic paradox" of believing P and non-P at the same time. Consider the dynamic paradox now. For the intentionalist account to be true, the agent cannot bring himself to believe that P, against evidence, in a straightforward way simply because he wants that P to be true. SD cannot be a direct and self-transparent strategy, because of the dynamic paradox. Hence if SD is to be intentional, it has to either indirect and/or non-transparent.

The indirectness has been proposed, exploiting time and bad memory, in such a way that S at $t^1$ can plan to lead herself to believe that P at time $t^2$ which now she knows it is false, as in the following example: If Clara wants to forget about a meeting fixed in two months time, so as to miss it without guilt, she can devise the stratagem to write it down on her diary at a wrong day. Given her poor memory, she is confident that in two months she will believe her own writing and forget the original date, so that she will believe the false and disbelieve the truth (Davidson, 1985; Mele, 1987, pp. 132–34; McLaughin, 1988, 1996; Bermudez, 2000).

But even if the example shows that it is conceivable to manipulate one's beliefs willfully, and cunningly create a false belief *ad hoc*, it does not show that this is a case, let alone a typical one, of SD, because in fact what S did was basically putting herself in the condition of believing P, which is false, in the usual rational way.[7] At time $t^2$ Clara will be justified in believing that P though P is false, given the evidence then available to her, so that she will not be in a state of SD, but rather in one of delusion.[8] If by contrast, Clara suddenly recollected what she had planned and done to deceive herself, the belief that P could not survive and the goal of peace of mind would definitely vanish. Indirectness is a self-defeating strategy for SD; let explore then the non-transparency option for making intentionality logically and conceptually possible. The non-transparency condition as a rule implies some reference to the unconscious, whether patterned after the Freudian notion, which may or may not lead to mind partition (Davidson 1985, Pears 1985, 1991). Leaving aside mind partition, which has been widely criticized, many scholars make use of a non-technical notion of unconscious, such as non-awareness, intrinsic opacity of cognitive operation, mental tropisms and so on (Gardner, 1983;

[7] That self-induced deception is not real SD is argued by McLaughlin (1996), while it is defended by Bermudez (2000).

[8] This is the argument made by Scott-Kakures (1996).

Talbott, 1995; Rorty, 1983, 1988; Barnes, 1998). Such explanations often sound as *ad hoc* accommodations with intentions which cannot in principle be acknowledged by the subject. For, there is a general methodological difficulty of non-transparent intentional accounts, namely the problem of SD ascription. Much as SDS cannot acknowledge SD's purpose as hers, SD can never be, and never is, self-ascribed in the present tense, because that would indeed be paradoxical, and no one could in fact acknowledge being self-deceived without exiting SD *ipso facto*. Therefore it happens that SD ascription is always made from outside without the possibility of being confirmed by SDS.[9] This very fact casts some doubt about the interpretation of SD as the subject's strategy. It is indeed an external observer, or a later self, who detects the false belief despite the contrary evidence, then find out the motivating wish, and the purpose behind SD. In a word, it is the observer who sees all the bits of a piece of practical reasoning in place: motivating wish, end and means; therefore, quite naturally, the observer is drawn to the conclusion of an intentional, though somehow unconscious, plan. Yet it is a plan which is in principle excluded that S can ever acknowledge in the present tense, and for which the observer lacks any clear and independent criterion of assessment (van Fraassen, 1988). The presence of a purpose and of a motive, supposedly evident to everyone, does not justify the inference of a strategy unconsciously devised by S. After all, the natural and social world displays a variety of seemingly purposive phenomena which are, in fact, unintended consequences of blind processes or of elsewhere directed actions. In a way, as professional observers, philosophers must be extra careful in order to avoid duplicating the illusions of SDS. Even if the teleological scheme is there, ready-to-use, familiar, well-embedded in every-day-life and common experience, we cannot just cash out its intuitive evidence eluding the methodological problem of outside ascription altogether. In order to retain the unconscious strategy account, a persuasive explanation of how the plan is carried out by a unified subject albeit non-transparently must yet be provided. In general, even the most persuasive versions of the intentionalist account, such as Fingarette's (1998), are obscure about what is the content of the self-deceptive intention: almost everyone excludes that it is the intention of deceiving oneself which would be puzzling indeed. But then: is it the intention to believe P which is knowingly false, or is it the intention to reduce one's

[9] The problematic ascription condition for SD is relatively overlooked in the literature, but see for example Johnson (1997, p. 104).

anxiety or improve one's image, and so on? The latter is definitely present and legitimately so; but the self-deceptive outcome, the soothing false belief, can hardly be seen as the direct result of that intention working in its usual way. (Hence the problem of explaining how that intention can work behind the back of the subject, so to speak, and the question whether this non-transparent work can be said "intentional" nonetheless). By contrast, the former, i.e., the intention to manipulate one's cognitive process in order to believe what one wishes, a) brings along the paradox and b) is simply imputed by the observer illegitimately, by applying the teleological scheme and by ascribing the apparent purpose to the agent. Even if the false belief is shown to be practically rational according to Bayesian rationality, this is not enough to prove the intentional strategic nature of SD processes (Talbott, 1995).

There is a point in favor of the intentional view, though, pointed out first by Talbott (1995). His defense of the intentionalist account refers to the lack of a satisfactory anti-intentional model for SD. He argues that if it were the case that a wish causally triggered a biasing process ending in a false belief, as anti-intentionalists maintain, there would be no limit to perceptual distortion for the immediate goal of maximizing pleasure and minimizing pain, with serious problem for the agent's long-run interests. For example, says Talbott, if I realize that the brakes of my car are not working well, that obviously worries and annoys me. But if I reacted to such worries simply by falsely coming to believe, as I wish, that my brakes are just fine, it would be very dangerous indeed. Instead, though it is a nuisance, I stop the car, and call up the garage, and patiently wait on the road until they come to pick me up, as it is reasonable to do in such cases. If *ex hypothesi,* however, SD is causally produced by a wish to reduce one's anxiety, by believing everything is fine, then why is it that, in the brake case, my mental processes do not take the first shortcut to pain minimization? If SD were the outcome of mental tropism for anxiety reduction, there would be no possibility of a different response in the brake failure case. This is why Talbott holds that we need an intentionalist account of SD, one which makes sense of SD limited scope in a fairly circumscribed area of individual life. Similarly Bermudes (2000) states that the selectivity of SD needs to be accounted and that causal explanations have so far no convincing answer. Yet the supposed deficiencies of causal explanation cannot prove that SD is an intentional strategy performed by a Bayesian agent.

## 2. The causal account

1. The anti-intentionalist view states that SD is a purely causal phenomenon where the operating cause is a motivational state, either a desire or an emotion, which activates cognitive biases impairing correct belief-formation; among the various causal interpretations of SD (Mele, 1987, 1997, 2001; Lazar, 1997, 1999), here I will mainly focus on Mele's, which is probably the most discussed in the last decade. He outlines a deflationary account of SD, which does away with all the puzzling aspects of the phenomenon, and explains the deceptive belief as caused by the interference of a wish with the usual way of lay hypothesis testing, manipulating the acceptance/rejection threshold for believing that P. Briefly, the every-day hypothesis testing theory (Friedrich, 1993; Trobe-Lieberman, 1996) says that our knowledge is generally oriented by the pragmatic need to minimize costly errors in belief-formation relative to resources required for acquiring and processing information. Individuals have different acceptance/rejection thresholds of confidence relative to the belief that p depending on the cost to the individual of a false acceptance or, conversely, of a false rejection. Motivations precisely interfere by manipulating the threshold, causing either to lower the acceptance threshold for believing that P or to heighten the rejection threshold for believing non-P; and this will result in a corresponding relaxation in the accuracy of data processing and evaluation, bringing the subject to falsely believe that p. In this way, there is no need to overcome any paradox, for the subject does not entertain two contrary beliefs, nor is necessary to imagine a person involved in a cunning manipulation of her mental states aimed at fooling herself. SD is indeed one species of motivated irrationality which exploits the normal everyday process of hypothesis testing and cognitive biases affecting all human cognition. In sum, for SD to be the case, in Mele's account is thus sufficient that:

1. the belief that p which S acquires is false;
2. S treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way;
3. this biased treatment is a nondeviant cause of S's acquiring the belief that p, and[10]
4. the body of data possessed by S at the time provides greater warrant for

[10] This condition is supposed to rule out that the deception is produced in someone other's than the subject.

non-p than for p. (Mele, 2001, pp. 50–51)

5. The attractiveness of Mele's account is easy to see: it is simple, non-mysterious, unified and backed on experimental psychology's model of lay hypothesis testing.[11] However many commentators doubt that Mele has indeed explained SD, instead of motivated beliefs in general, such as wishful thinking, or even unmotivated biased beliefs, such as delusion (Audi,1988; Bermudez, 2000; Neklin, 2002). Mele's sidestepping of the paradox, by positing a one-belief explanation, indeed creates a trouble of that kind. For SD to be the case, actually it is not sufficient (1) that p be false, and (2) that the relevant data are treated in a biased way, given (4) that the data possessed by S provides greater warrant for non-p that for p. In Mele's own previous description (1987), SD is believing in the teeth of evidence. So, on the one hand, the evidence available to S must provide not just greater, but significantly greater (though not conclusive) warrant for non-p than for p, so that any independent observer would easily conclude that non-p. For if the evidence is ambiguous, the subject may conclude that p, which corresponds to her motivation and is false, but still is held in a rationally justified way. On the other, the counter-evidence must be appraised by the subject, since it is precisely that appraisal which activates the wish, and sets off the SD process. If there is no such appraisal of the contrary data, as maintained by Mele, that implies that the motivationally relevant counter-evidence is automatically shut off S's awareness and stored in some non conscious mind module; but then, the relevant evidence is not available to S –-contrary to what is stated in (4); and, in the absence of contrary evidence, her belief-formation pattern works correctly even if it ends up falsely believing that P. If by contrast, the belief-formation pattern is irrational as typical of SD, then the appraisal of the counter-evidence is a necessary condition. Such appraisal does not need to produce the corresponding belief non-P (Greenwald, 1988, p. 127), but it should lead S, if she is a normally rational person, as SD implies, *considering*, *suspecting* that non-p is the case (Michel & Newen, 2010). If S is blind to the evidence, and comes to believe that p in the usual way, then when an external observer points out the missing evidence to her, she should be in the position to accept the criticism and revise her belief, because indeed her mistake was due to the lack of relevant evidence;

[11] On the general pragmatic model of hypothesis testing see: J.Klayman and Young-Won Ha (1987). On the explanation of SD by this general model see James Friedrich (1993).

actually, one of the phenomenological feature of SD lies precisely in that S defends her deceptive beliefs against criticisms and does so in a reason-like style providing arguments, no matter how faulty, supporting her belief and explaining away the counter evidence (Forrester, 2002); while a false non-deceptive belief, causally produced by biases, is usually willingly revised by subjects. So the deflationary move of one-belief explanation risks to loose its object, SD, and what explains instead is a general kind of motivated irrationality. But for SD to be the case, we can well dispense with the two contradictory beliefs (and Mele is right in that), but we cannot dispense with the appraisal of the negative evidence which, moreover, makes sense of another phenomenological feature of SD, namely the internal tension of the subject which characterizes most cases, if not the whole of SD (Audi, 1983, 1988).

6. Another problem that Mele's simple and unified account has to face is the selectivity problem seen above. If SD is the process and the resulting state by which our desires causally distort cognition by activating biases, how come that not *all* desires *always* become operative in that sense, and that most of the time we come to hold rational beliefs? The problem has elicited the following answer by Mele: think of the case of Gideon, a CIA agent accused by treason. While both his staff and his parents share the desire that he is innocent, when confronted with the body of evidence, his staff comes to be convinced that he is guilty, but his parents retain the deceptive belief that he is innocent. SD hence applied only to the parents' belief. Mele's explanation of the difference is that the cost of falsely believing Gideon innocent is higher for the intelligence agents than for his parents. That is because for the staff, the desire of his innocence is trumped by the desire of not being betrayed. However this explanation has hardly explained how SD works selectively: it seems clear that having the desire that p is not sufficient to biasing data treatment; but then we must have a theory which specifies *which desires* in S's motivational set may become operative for biasing, and in *which situations.* But then the simple and unitary explanation, referring only to motivation and causal biasing, needs to get much richer and more complicated, in a way that Mele clearly wants to avoid.

In general the reference to the lay model of hypothesis testing, which apparently provided experimental backup to Mele's view, being a general explanation of normal reasoning in normal circumstances, can backfire on his account. For the model says that human cognition is always pragmatically

rather than epistemically oriented and likewise open to pervasive biases and systematic mistakes. But then a) motivations normally intertwine with cognitive processes, and b) biases are normally ubiquitous. How can we specifically detect SD in such a cognitive background? If despite pervasiveness of biases and motivation interference, on the whole, we are responsive to evidence and come to hold beliefs which are mostly true, then we cannot explain the specificity of SD via our general cognitive vulnerability.

## 3. Mental trap or cunningly planned?

3.1. A purely causal story of SD discounts the propositional nature of SD doxastic process. A recent work by Michel and Newen (2010) refers to the experiments by Wentura and Greve (2003, 2005) on how subjects adapt trait-definition for self-immunization purposes. Subjects who, *ex ante,* have thought of themselves as cultivated and, specifically, knowledgeable in history, and, in the context of the experiment, have failed history test, immediately processed the negative result by adapting the "criterial evidence" required to define someone "cultivated". Adapting the previous belief that "knowledge of history is a necessary component for a cultivated person", or discounting the value of the test for real historical knowledge, subjects managed to defend the belief that they were cultivated, in the teeth of contrary evidence. That such stories were self-deceptive is proved by the fact that the subjects, who were tested as normally rational and evidence-sensitive in general, applied standards of evaluation and reasoning to themselves different from those usually applied in general and specifically to other people. Yet, and this is the aspect I want to stress, their stories were construed in an argument-like fashion and presented in a seemingly coherent set of propositions. In other words, those self-deceptive stories did not look like a causal result of biases operating behind the subjects' back, but like the result of an intentional effort not at deceiving themselves but at finding a way out of self-embarrassment. The subjects' reasoning was twisted, no doubt, and suspicious, given the unwarranted shift in the "criterial evidence" for being cultivated, nevertheless it responded to usual constraints on reasoning, for example providing an account of the negative evidence, no matter if by means of *ad hoc* explanations, and making use of arguments, no matter how unsound. Michel and Newen conclude that SDS displays dual rationality and that what constitutes self-deceptive reports is a

quasi-rationality working in an automatic, pre-reflexive, hence non transparent mode to the subject.

Drawing from this work as well as from daily experience, it seems that the dynamic of SD can hardly be accounted as a mental event induced by a motivational state that switches on cognitive biases which, in turn, non-deviantly cause the false belief that P. It seems that there is a lot that subjects do and do knowingly, and up to a point openly and legitimately, which grounds the reasoning towards the belief that P though such reasoning is typically faulty (Forrester, 2002). Yet that the process is done by the subject and not merely happens to her, does not imply that it is actually aimed at procuring the self-deceptive belief. If the anti-intentionalist account, on the one hand, cannot distinguish different varieties of motivated beliefs, and, on the other, cannot explain why desires sometimes lead to accurate response and sometimes to SD, the intentionalist account stumbles on paradoxical view. The standoff between intentionalists and causalists is partly produced by a lack of clarity about what the intention should be for SD to be intentional. Most prominently, the distinction between *intentionality of process* and *intentionality of outcome* is blurred. For the outcome to be unintended it is not necessary that the process is likewise unintended and causal. Nor do we need to think of an unconscious mind as the agent, inaccessible to the conscious ego, to account for the production of a deceptive belief which cannot be self-ascribed in the present tense. The best solution must account both the intentional steps and the unintentional deceptive belief which results from the process, and I propose that the model of invisible hand be such a candidate. An invisible hand explanation for SD does away with the paradoxical idea of lying to oneself; and yet it can account the purposive appearance of SD without recourse to a deceptive plan which would not sit comfortably with the impossibility of ascription in the present tense; moreover, it can also capture the distinctiveness and selectivity of SD which are lost in a purely causal deflationary account. In other words, it seems to me that if SD is to be accounted a) as a genuine and ordinary phenomenon; b) as a non-mysterious, nor paradoxical process; c) as a distinct specimen of motivated irrationality, then *it cannot be:* a) intentional pretense; b) an intentional, though partly unconscious, plan; c) a purely causal happening. In order to accommodate the apparent purposiveness, the non-intentionality of the outcome and the selectivity of the process of deceptive belief formation, SD must be conceived

along the invisible hand model: as an intentional doing otherwise directed, whose deceptive outcome is unintended, though serves an aim of the subject.

3.2. What the subject does when she appraises of threatening evidence for the belief that P may be done in a pre-attentive mode, and may not require full awareness, but it is her doing. The wish that P and the desire to defend the belief that P are legitimately there, can even be acknowledged by S, and need not be the causal trigger of SD process. Actually the consequent search for an explanation which can accommodate P with the negative evidence is intentionally taken up by S and, I would add, legitimately so. So far, no irrational move has yet been made. However, once the process of thinking and of considering evidence starts, S has to make interpretative choices, given that, by definition, the evidence available, though clearly unbalanced in favor of non-P, is not conclusive and does not compel her to believe that non-P. Again, this is quite a normal cognitive situation, and it is also quite a normal fact that those choices are often influenced by extra-epistemic facts: heuristics, past experiences, proximity, salience of various kinds, aesthetic values, asymmetry between the evidence required believing something new and to disbelieve something taken for granted. Some of these extra-epistemic elements are what cognitive psychology has called cold biases, and has detected as intrinsically winded up with intelligent thinking. In this case, however, among the extra-epistemic factors, there is especially the wish that P.

How is the wish that P working on the cognitive process that S has started in order to assess the evidence against P, and possibly to defend the belief that P? Three options have been put forward in the literature: a) the wish works exactly like any other desire (short of the confusion between reality and beliefs), providing reasons for action to the subject who then devises an intentional strategy aimed at securing the goal of believing P (Gardner ,1983); b) the wish to believe that P is reflected in the preference ranking of the subject, who proceeds to intentional biasing in order to secure the belief that P (Talbott, 1995); c) the wish causally triggers the biasing ending up with the belief that P (Mele, 2001). None seems to me correct. Firstly, the wish does not work like a normal strong desire providing reason for action aimed at states of the world, precisely because changing the state of the world is beyond the scope of SD (we'll come back on this shortly). That is why, instead of acting, the subject lingers in thinking. Secondly, I would describe the influence of the anxious wish on S's thoughts neither as a motive for intentionally biasing, nor as a mere

cause for blindly biasing. It seems to me that in the process of reflection, the wish intervenes when interpretative choices are to be made, much in the same way as a theoretical hypothesis intervenes in scientific research, orienting the analysis in a certain direction, raising certain questions and discarding others, searching to the left and not to the right. This intervention seems both intentional and, in a way, legitimate, given that contemporary epistemology has amply acknowledged that facts do not speak for themselves and that theoretical frameworks are necessary for providing meaningful accounts (Sultana, 2006). Experimental psychology confirms that in daily reasoning, subjects tend to be guided less by epistemic norms than by heuristics. I think that the anxious wish works precisely as a pragmatic influence, selecting the focal error to be avoided, orienting the direction of thinking, the search and assessment of facts for reaching a judgment. In this influence, I see neither a self-deceptive intent, nor a self-deceptive event at work yet. The wish works as a pre-theoretical and extra-epistemic pragmatic selector; and the fact that in this case the selector is "motivated" is not a distinctive element either, given that very often intuitions orienting scientific research are motivated as well. In this process, then, cold biases can possibly kick in, but again, such interference is not specific to SD, being rather the normal condition of human intelligent thinking.

So what does it make for a difference, if at all, in cases that we label SD? I can think of two main differences. The first is that when S has found an explanation realigning the unpalatable facts with the desired reality, she sits on it, no matter how unlikely such possibility appears to anybody else. In other words, as soon as S is capable of explaining away the evidence against P, she stops her search and reasoning. And this sudden stop is not typical of any "cold" inquiry, though influenced by pre-theoretical hypothesis and extra-epistemic values. In cold cases of HT, despite the pragmatic orientation, S is more cautious and the threshold of evidence deemed necessary to believe P is considerably higher. SDS, by contrast, has a suspiciously low threshold of required evidence, as Mele has well underlined, so she stops as soon as she finds the way to go on believing that P, no matter how implausibly. This is precisely an (epistemically) irrational move. Is it causally induced or intentionally done? In a way, it is something in between: it is the agent who stops there, and she knows that she stops, and this is done intentionally, even though without a specific deliberate choice; yet the general meaning of this move escapes her, as long as it is possible for her to believe that P. In other words, it escapes her that her conclusion is unwarranted, and that her

reasoning has been faulty. The second difference is that the non-transparency of the SD process is a specifically thematic one. It is not simply that we do not master our cognitive processes and that cold biases are pervasive and beyond our control; that, again, is common to any cognitive enterprise and in no way can single out, let alone explain, SD. The non-transparency of SD is a special kind of overall opacity possibly caused by the strong emotional state of the subject, which somehow impairs her cognitive lucidity about the whole process and its outcome. But it is important to grasp how this impairment works, because it is not like when a sudden fright blocks our perception and distorts our cognition directly. In SD cases, by contrast, S does not experience herself as a victim of an emotional grip because any single step in the production of SD is both intentional and transparent, under a piecemeal description. The cognitive opacity concerns the overall process whose meaning escapes S and about which her usual critical appraisal seems to be blocked. In other words, the emotional grip induces a general relaxation of usual epistemic standards so that S does not detect the cognitive inadequacy of the cover story, and is contented to have devised a support for her belief that P.

3.3. Let see now how this account can sort out the selectivity problem. Both Talbott (1995) and Bermudez (2000), who have raised this issue against the causal account, seem to think that the intentional view preempts such a problem, given that the selection is directly made by the intentional agent wanting to bring about the belief that p. However this solution seems to presuppose that the crucial intention for SD is precisely that of deceiving oneself, an intention verging on the paradox which I have excluded to be part of the invisible hand account. In my perspective, the selectivity problem must be differently addressed. Robert Jervis (1976) points out the expected utility of the information as the reason for different degrees of accuracy in testing data and forming a proper belief. If the cost for inaccuracy is high, it is likely that the agent will adopt a vigilant attitude, while if the cost is low, accuracy can be dispensed of. This implies that if the cost for inaccuracy is low, the interference of a desire on cognition has more probability to happen than when the cost is high: and this fits with the case of the brake failure. But then Jervis also acknowledges that costs and incentives are not the whole story; selective vigilance or inaccuracy correlate as well with the level of anxiety and stress concerning the evidence. Low and high anxiety would typically induce less accuracy than medium level of stress. But while low anxiety leads the agent to

rely on routines and traditional pattern of conduct, high anxiety and stress tend to engender "defensive avoidance" that is a blocking of the negative information and reliance on a false soothing belief, i.e., SD. The two stories for the variance of vigilance/inaccuracy in evidence processing can be interestingly combined: if the cost for inaccuracy is low and the level of stress likewise low, then habitual response and traditional pattern follows. If the costs are high and the level of stress medium, such as in the brake-failure case, then accuracy is higher and optimal response follows. If the anxiety and stress induced by certain evidence are very high, and if the agent perceives the situation as beyond his or her control, then we have typical circumstances for SD to take place: the costs of inaccuracy are irrelevant since the agent cannot change the state of the world while the deceptive belief will relieve anxiety, at least in the short term. When the stress level is very high, and the costs of inaccuracy are also high, what follows is a variable of the psychological conditions of the agent, and of her capacity to stand and to respond rationally to stressful stimuli.

In this way the selectivity of SD is accounted by low cost of accuracy in data processing and strong emotional load in the perceived discrepancy between evidence and desire. Such explanation excludes a purely causal account of SD for it implies that the subject not only appraises the negative evidence and detects its potential threat, but also senses whether vigilance is required to overcome the threat or not. Meanwhile also the desire that P at the origin of SD process can be similarly specified: it is emotionally loaded because that P be and be believed true is often crucial for the subject, and beyond his control.[12] The wish that P often concerns mortal questions, either in a literal or in a symbolic sense. By mortal questions I mean matters which bear a fundamental and constitutive relationship with the self.[13] A brief survey of all examples used to illustrate SD points out that matters of SD are usually death, love and self-esteem or self-respect, that is, matters which are crucial for one's balance and well-being. Other cases look less tragic: often we re-describe unwelcome

[12] That the desire originating SD must be "anxious" is stated by Pears (1985) and Johnston (1988), denied by Mele (2001), and discussed by Michel-Newnen (2010), concluding that it is not necessary.

[13] The expression comes from Nagel 1979. However I would stress that the momentous nature of such questions derive from the relationship the subject sees between them and herself, more than in the essential features of certain problems. Though most of examples for SD are indeed of such momentous nature, not every scholars share the view that SD has to do with mortal question: see, for example, Rorty (1996, pp. 75-89), where she puts forward a sort of naturalistic explanation of SD as a sort of functional device to cope with complex natural and social environment.

truths about ourselves in a way to realign the negative evidence – failures and misconduct of various kinds – to the positive self-image we harbor and cherish in our bosoms. In the reduction of cognitive dissonance between evidence and self-image the costs of accuracy are also low, because failures have taken place already, and a diagnostic self-reflection would only make people feel depressed, guilty and powerless, while a deceptive positive image can enhance a more energetic or adaptive response. How distressing is the negative evidence can vary; but whether it is a case of mortal question or of a more familiar and daily disappointment, if the costs for inaccuracy are low, the SD response is likely to happened. When relatively trivial negative evidence bothers the self, as for the fox with the grapes, the deceptive belief which reduces the cognitive dissonance is generally more stable, because it is less likely to be undermined by further negative evidence coming in. By contrast, when mortal questions are at issue, SD provides only a palliative treatment, and the subject is always, though within lapses of time, haunted by the evidence explained away by the cover story, but never finally buried, because SD can make one believe that P, but cannot make P true. Thus the subject believes that P, but is constantly presented with a reality which makes P very unlikely because the disquieting evidence does not stop to come in. In other words, the very nature of the wish that P excludes that P be the goal of an intentional strategy aimed at its fulfillment, precisely because securing P is beyond the control and possibility of the subject, whether it is a mortal question or a more mundane failure. We can thus set apart desires which put in motion a self-deceptive process, from other emotional demands which engender either rationally adequate responses, or other, less sharp, forms of motivated irrationality. The candidate for SD must be not only a self-serving, emotionally loaded desire, but also one that S cannot fulfill by usual rational action. When this kind of desires is met with contrary evidence which, though not conclusive, would lead a rational person to believe non-P, then the circumstances for SD to take place obtain, circumstances which should enter in any account of SD, and likewise supplement conceptual analysis for SD to correspond to our distinctive intuitions.

Once we have singled out the appropriate kind of wishes as points of departure of the deceptive process, we need not suppose that they work as a causal triggers of biasing belief-formation, for we have seen that, from one perspective, SD is all of the subject's doing: indeed

a) S starts thinking over the disquieting facts;

b) S, selectively retrieving, imagining, piecing together, comes up with an explanation of why P is the case, despite the contrary evidence;
c) S hangs on the cover story and believes it, no matter how implausible;
d) S accepts the (false) belief that P and disposes of the very idea that non-P;
e) as a result, anxiety and worries are dispelled – for the time being – via a manipulation of one's doxastic states.

Yet, from another perspective, S neither plans her deception nor directly performs her beliefs' manipulation. She has no sense of what she is doing putting all steps together. With the exception of (c) and partially of (d), each move is epistemically legitimate, and all are intentionally taken, though not necessarily in full awareness and never considered in a sequence as a comprehensive strategy. It is only when they are all pieced together by an external observer that a strategy can be seen, a strategy aimed at the goal of reducing anxiety, via the pacifying belief that P. But this strategy has never been the subject's, though fulfilling her practical goal of finding some peace of mind. It is the unintended outcome of different steps elsewhere directed, actually directed at reconsidering evidence and forming a true judgment, and only one of which – move (c) – is specifically faulty corresponding to the quasi-rationality highlighted by Michel and Newen. Such non-transparent quasi-rational mode prevents S from having a comprehensive view, let alone a critical one, of the whole process. In this sense, she is a victim and not an agent of her SD. And from this viewpoint, SD is unintentional, brought about by a joint effect of single intentional moves, plus the causal interference of the emotion inducing a lapse of proper rationality so that the subject uncritically endorses the cover story and candidly comes to hold the false belief. The invisible hand account reconciles the apparent purposiveness of SD with the impossibility of conceiving it as a strategic plan of the subject. That has been disposed by the circumstances for SD. Since the agent cannot dispel her worries by engaging in action aimed at changing the state of the world, she cannot likewise intentionally engage in SD, which corresponds to her second best preferences, ie. to believe that P, contrary to available evidence. SD cannot be an intentional strategy not only because it would imply a paradox, but also because it can never be self-ascribed in the present tense. To be sure, peace of mind can be reached by a false belief; but, even assuming that one can make oneself believe a false belief at will, no one could devise that as a strategy for reaching peace of mind, because, from the agent's viewpoint in that very moment, what does the

job of relaxing her anxiety is that P is a true state of the world, not the belief that P, no matter what. The exchange between unfavorable states of the world and benign beliefs cannot be an intentional trade-off, because it would precisely make the desired peace of mind impossible, being S normally rational and constrained by responsiveness to evidence. So unless the false soothing belief is brought about by intentional moves but not aimed at believing against the evidence, the subject cannot candidly endorse that P and SD would be self-defeating.

At a later time, S may acknowledge her previous SD, and she usually feels shame and blames herself at having been such a fool, though at the time she could not help it. Can we also blame S for being self-deceived? As I see the problem, the answer depends on whether S may avoid ending up with unjustified and self-serving beliefs. The avoidance of SD cannot be helped by exhortation, or self-exhortation, because the process is not precisely under S control. But if not directly, one can learn how to control one's actions and beliefs indirectly. Moral psychology has singled out at least two forms of indirect control, just in order to bypass akrasia: character-building, via moral learning and discipline (Aristotle; Ainslie 2000), and pre-commitment (Elster, 1980). Both requires that S feels shame and regret at her previous SD and is willing to do what is necessary to avoid falling prey. Moral learning implies to detect the circumstances favorable to SD and adopt a vigilant attitude, having fortified one's character with moral discipline. It may not suffice though; pre-commitment, the strategy to create some constraint on one's options at $t^1$, under condition of cognitive lucidity, so as to avoid at $t^2$, under emotional pressure, being prey of temptation one knows it is difficult to resist, may be more promising. S can trust oneself to a referee, so to speak, concerning one's motivated hypothesis. Reversing what usually happens in SD cases, when the self-deceptive belief is often supported by a charitable community (Rorty, 1996; Salomon, 1996), the subject should confer her friend(s) the authority of referee(s) in case of beliefs held in the teeth of evidence. Such authorization is important. For, in the first place, the friends of the prospective SDS should avoid the self-appointed role of guardians, with its implicit self-righteousness, and, in the second place, the agent ought to take responsibility for their intervention in order to subscribe his (pre) commitment against SD. Conversely, just because SD is avoided through the assistance of a friend acting as a referee for one's belief, the agent can take credit of SD avoidance only with an explicit authorizing agreement, made *ex ante*, under condition of

cognitive lucidity. Thus the agent becomes properly responsible of her SD in case she dismisses the referee's advice, or of her avoidance.

## REFERENCES

Aislie, G. (2001). *Breakdown of the Will*. Cambridge: Cambridge University Press.

Ames, R.T., & Dissanayake, W. (Eds.) (1996). *Self and Deception: A Cross-Cultural Philosophical Enquiry*. Albany: State University of New York Press.

Aristotle (1988). *Nicomachean Ethics*. (tr. by D. Ross). Oxford: Oxford University Press.

Audi, R. (1982). Self-Deception, Action and Will. *Erkenntnis*, *18*(2), 133–158.

Audi, R. (1988), Self-Deception, Rationalization, and Reasons for Acting. In B.P. McLaughlin & A.O. Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 1988, 92–20.

Audi, R. (1989). Self-Deception and Practical Reasoning. *Canadian Journal of Philosophy*, *19*(2), 247–266.

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

Barnes, A. (1998). *Seeing Through Self-Deception*. Cambridge: Cambridge University Press.

Bermudez, J.L. (2000). Self-Deception, Intention and Contradictory Beliefs. *Analysis*, *60*(4), 309–319.

Davidson, D. (1982). Paradoxes of Irrationality. In R. Wollheim & J. Hopkins (Eds.), *Philosophical Essays on Freud*. Cambridge: Cambridge University Press, 289–305.

Davidson, D. (1985). Deception and Division. In E. LePore & B. McLaughlin (Eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell, 138–148.

Demos, R. (1960). Lying to Oneself. *The Journal of Philosophy*, *57*(18), 588–595.

Dupuy, J.P. (Ed.) (1998). *Self-Deception and Paradoxes of Rationality*. Stanford: CSLI Publications.

Elster, J., (1980). *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

Elster, J. (1983a). *Explaning Technical Change*. Cambridge: Cambridge University Press.

Elster, J. (1983b). *Sour Grapes. Essay on Rationality and Irrationality*. Cambridge: Cambridge University Press.

Elster, J. (1985). Deception and Self-Deception in Stendhal: Some Sartrian Themes. In J. Elster (Ed.), *The Multiple Self*. Cambridge: Cambridge University Press, 93–113.

Fingerette, H. (1969). *Self-Deception*. London: Routledge & Kegan Paul.

Fingarette, H. (1998). Self-Deception Needs No Explaining. *Philosophical Quarterly*, *48*(192), 289–301.

Foss, J. (1980). Rethinking Self-Deception. *American Philosophical Quarterly*, *17*(3), 237–243.

Forrester, M. (2002). Self-Deception and Valuing Truth. *American Philosophical Quarterly*, *39*(1), 31–47.

Friedrich, J. (1993). Primary Error Detection and Minimization (PEDMIN) Strategies and Social Cognition. A Reinterpretation of Confirmation Bias Phenomenon. *Psychological Review, 100*(2), 298–319.

Funkhauser, E. (2005). Do the Self-Deceived Got What They Want *Pacific Philosophical Quarterly, 86*(3), 295–312.

Gardner, S. (1983). *Irrationality and the Philosophy of Psychoanalysis*. Cambridge: Cambridge University Press.

Gergen, K.J. (1985). The Ethnopsychology of SD. In M.W. Martin (Ed.), *Self-Deception and Self-Understanding*. Lawrence: University Press of Kansas, 228–243.

Gilovich, T. (1991). *How Do We Know What Isn't So?*. New York: The Free Press.

Gur R.C. & H.A. Sackeim (1979). Self-Deception: A concept in search of a phenomenon. *Journal of Personality and Social Psychology*, *37*(2), 147-169.

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

Haight, M.R. (1985). Tales from a Black Box. In M.W. Martin (Ed.), *Self-Deception and Self-Understanding*: Lawrence: University Press of Kansas, 244–260.

Hamlyn, D.W. (1971). Self-Deception. *The Aristotelian Society: Supplementary Volume*, *45*, 45–60.

Jervis, R. (1976). *Perception and Misperception in International Politics*. Princeton: Princeton University Press.

Johnson, E.A. (1997). Real Ascription of Self-Deception are Fallible Moral Judgements. *Behavioral and Brain Sciences*, 20(1), 104.

Johnston, M. (1988). Self-Deception and the Nature of the Mind. In B. McLaughlin & A.O. Rorty (Eds.), *Perspective on Self-Deception*. Berkeley: University of California, 63–91.

Klayman, J., & Young-Won, H. (1987). Confirmation, Disconfirmation and Information in Hypothesis Testing. *Psychological Review*, 94(2), 211–228.

Kurzban, R. (2010). *Why Everyone (Else) is a Hypocrite. Evolution and the Modular Mind*. Princeton: University Press.

Lazar, A. (1997). Self-Deception and the Desire to Believe. *The Behavioral and Brain Sciences*, *20*(1), 119–120.

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

Martin, C. (Ed.) (2009). *The Philosophy of Deception*. Oxford: Oxford University Press.

Martin, M.W. (Ed.) (1985). *Self-Deception and Self-Understanding: New Essays in Philosophy and Psychology*. Lawrence: University Press of Kansas.

McLaughlin, B.P. (1988a). Mele's Irrationality: A Commentary. *Philosophical Psychology*, *1*(2), 189–200.

McLaughlin, B. (1988b). Exploring the Possibility of Self Deception. In B.P. McLaughlin & A.O. Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 1988, 29–62.

McLaughlin, B.P. (1996). On the Very Possibility of Self-Deception. In R.T. Ames & W. Dissanayake (Eds.), *Self and Deception*. Albany: State of New York Press.

McLaughlin, B. P., & Rorty, A. O. (Eds.) (1988). *Perspectives on Self-Deception*. Berkeley: University of California Press.

Mele, A. (1987). *Irrationality: An Essay on Akasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press.

Mele, A. (1997). Real Self-Deception. *Behavioral and Brain Sciences*, *20*, 9–102.

Mele, A., (1999). Twisted self Deception. *Philosophical Psychology*, *12*(2), 17–137.

Mele, A., (2002). *Self-Deception Unmasked*. Princeton: Princeton University Press.

Mele, A. (2009 a). Have I Unmasked Self-Deception or Am I Self Deceived?. In C. Martin (Ed.), *The Philosophy of Deception*. Oxford: Oxford University Press, 260–276.

Mele, A. (2009b). Delusional Confabulation and Self-Deception. In W. Hirstein (Ed.), *Confabulation. View from Neuroscience. Psychiatry, Psychology and Philosophy*. Oxford: Oxford University Press, 139–158.

Michel, C., & Newen, A. (2010). Self-Deception as Pseudo-Rational Regulation of Belief. *Consciousness and Cognition*, *19*(3), 731–744.

Nagel, T. (1979). *Mortal Questions*. Cambridge: Cambridge University Press.

Nelkin, D. (2002). Self-Deception, Motivation and the Desire to Believe. *Pacific Philosophical Quaterly*, *83*, 384–406.

Nozick, R. (1974). *Anarchy, State and Utopia*. Cambridge: Harvard University Press.

Nozick, R. (1977). On Austrian Methosology, *Synthese*, *36*, 353–392.

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

Pears, D. (1985). The Goals and Strategies of Self-Deception. In J. Elster (Ed.), *The Multiple Self*. Cambridge: Cambridge University Press, 59–77.

Pears, D. (1991). Self-Decetive Belief Formation. *Synthese*, *89*(3), 393–405.

Piattelli-Palmarini, M. (1994). *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York: John Wiley & Son.

Rey, G. (1988). Toward a Computational Account of Akrasia and Self-Deception. In B. McLaughlin & A.O. Rorty (Eds.), *Perspective on Self-Deception*. Berkeley: University of California, 264–296.

Rorty, A.O. (1988). The Deceptive Self: Layers and Loirs. In B. McLaughlin & A.O. Rorty (Eds.), *Perspective on Self-Deception*. Berkeley: University of California, 11–28.

Rorty, A.O. (1994). User-Friendly Self-Deception. *Philosophy*, *69*(268), 211–228. Reprinted in R.T. Ames & W. Dissanayake (Eds.), *Self and Deception: A Cross-Cultural Philosophical Enquiry*. Albany: State University of New York Press, 75–89.

Sahdra, B., & Thagard, P. (2003). Self-Deception and Emotional Coherence. *Minds and Machines*, *13*(2), 213–231.

Salomon, R.C. (1996). Self, Deception and Self-Deception in Philosophy. In R.T. Ames & W. Dissanayake (Eds.), *Self and Deception: A Cross-Cultural Philosophical Enquiry*. Albany: State University of New York Press, 91–121.

Sartre, J.P. (1956). *Being and Nothingness*. (tr. by H.E. Barnes). New York: Philosophical Library.

Scott-Kakures, D. (1996). Self-Deception and Internal Irrationality. *Philosophy and Phenomenological Research*, *56*(1), 31–56.

Scott-Kakures, D. (2002). At Permanent risk: Reasoning and Self-Knowledge in Self-Deception. *Philosophy and Phenomenological Research*, *65*(3), 576–603.

Siegler, F.A. (1968). An Analysis of Self-Deception. *Nous*, *2*, 147–164.

Sultana, M. (2006). *Self-Deception and Akrasia. A Conceptual Analysis*. Roma: Editrice Pontificia Università Gregoriana.

Szabados, B. (1974). The Morality of Self-Deception. *Dialogue*, *13*, 25–34.

Szabados, B. (1985). The Self, Its Passions and Self-Deception. In M.W. Martin (Ed.), *Self-Deception and Self-Understanding*. Lawrence: University Press of Kansas, 143–168.

Szabò-Gendler, T. (2007). Self-Deception as Pretense. *Philosophical Perspectives*, *21*(1), 231–258.

Talbott, W.J. (1995). Intentional SD in a Single, Coherent Self. *Philosophy and Phenomenological Research*, *55*, 27–74.

Torey, Z., (1999). *The Crucible of Consciousness. An Integrated Theory of Mind and Brain*. Cambridge (Ma): MIT Press.

Trobe, Y., & Liberman, A., (1996). Social Hypothesis Testing. Cognitive and Motivational Mechanism. In E. Higgins & A. Kruglansky (Eds.), *Social Psychology: Handbook of Basic Principles*. New York: Guildof Press, 239–270.

Vaillant, R. (1993). *The Wisdom of the Ego*. Cambridge: Harvard University Press.

Van Fraassen, B. (1988). The Peculiar Effect of Love and Desire. In B.P. McLaughlin & A.O. Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 123–156.

Velleman, D., (2005). The Self as a Narrator. In J. Christman & J. Andersen (Eds.), *Autonomy and the Challenges of Liberalism*. Cambridge: Cambridge University Press, 56–76.

Wentura, D., & Greve, W. (2003). Who Want To Be...Erudite? Everyone! Evidence for Automatic Adaptation Trait Definition. *Social Cognition*, *22*(1), 30–53.

Wentura, D., & Greve, W. (2005). Evidence for Self-Defensive Processes by Using a Sentence Priming Task. *Self and Identity*, *4*(3), 193–211.