Against the Deflationary Account of Self-Deception*

José Eduardo Porcher†
jeporcher@ufrgs.br

ABSTRACT

Self-deception poses serious difficulties for belief attribution because the behavior of the self-deceived is deeply conflicted: some of it supports the attribution of a certain belief, while some of it supports the contrary attribution. Theorists have resorted either to attributing both beliefs to the self-deceived, or to postulating an unconscious belief coupled with another kind of cognitive attitude. On the other hand, *deflationary* accounts of self-deception have attempted a more parsimonious solution: attributing only one, false belief to the subject. My aim in this paper is to critically examine this strategy and, subsequently, to suggest that its failure gives support to the neglected view that the self-deceived are not accurately describable as believing either of the relevant propositions.

Introduction

Alfred Mele[1] and others have rightly eschewed the literal understanding of "self-deception," calling attention to the phenomenon we want to explain rather than the word we use to refer to it. This liberates us from having to prove that self-deception, an obviously widespread phenomenon, is possible. Regarding the mental state in which the literally self-deceived would be in, the so-called static puzzle results from realizing that whereas in interpersonal

* The author would like to thank Michael Losonsky, Eric Schwitzgebel, Neil Van Leeuwen and an anonymous referee for helpful comments.

† Federal University of Rio Grande do Sul, Brazil.

[1] All references to his work will be to Mele 2001. Much of the material for that book comes from Mele 1997a, 1997b, which develop ideas that are already present in Mele 1987a. For Mele's answers to recent critics, see Mele 2009.

deception someone believes that p and causes the belief that not-p in someone else, in an intrapersonal analogue of interpersonal deception, these two beliefs would have to coexist simultaneously. Why is this puzzling? Mainly because most accounts of belief take it to be constitutive of belief that the content of what one believes is what guides one's thoughts and actions, so that if we were to attribute simultaneous, contradictory beliefs to a subject, such an attribution would not have any explanatory or predictive power. Because literal self-deception necessarily involves this kind of attribution, literalists such as David Pears have found that «self-deception is an irritating concept. Its supposed denotation is far from clear and, if its connotation is taken literally, it cannot really have any denotation» (1984, p. 25). Which is to say that, apart from the very difficulty of arriving at a consensual definition, the very word "self-deception" carries with it an air of impossibility if we take it to mean exactly what it seems to mean.[2] Adherence to the literal reading has resulted in various strategies to solve the resulting puzzles. The key characteristic they share is that all of them splinter the mind somehow, literally, into separate, fully rational and autonomous subagents (Pears 1984), or functionally, into separate, independent compartments (Davidson 1982, 1985).[3]

Those who have distanced themselves from the literal interpretation of "self-deception" have felt that the pull toward the attribution of simultaneous contradictory beliefs is still present, and so, that the puzzle still demands an answer. This is because the behavior of the self-deceived is (at least in some cases) deeply conflicted: many times the verbal behavior of the self-deceived will indicate that they believe that p and their nonverbal behavior will indicate that they believe that not-p. Worse yet, in some cases the nonverbal behavior as a whole will be inconsistent, so that the self-deceived will sometimes act and react in ways that indicate that they believe that p, and other times in ways that indicate that they believe that not-p. There has been one main strategy to account for this fact while withholding the attribution of contradictory beliefs.

[2] Literal self-deception also engenders another, dynamic puzzle, which results from modeling self-deception on intentional deception. This way, the self-deceived would have to engage in an impossible project: to intend and, at the same time, to hide one's intention from oneself. For more detail on the static and dynamic puzzles and some of the attempted solutions, see Mele 2001, pp. 3-24.

[3] As my aim is to critically assess the deflationist position, I will not concern myself here with the problems raised by postulating mental division. But see Johnston (1988) for criticism of Pears (1984), and Heil (1993) for criticism of Davidson (1982, 1985).

The key characteristics its varieties share are, on the one hand, the attribution of an unconscious belief in the undesirable proposition that the evidence favors and that motivates the self-deception; and, on the other hand, the attribution of another kind of cognitive attitude toward the content of the false or unwarranted proposition that the self-deception is about. Some have maintained that the attitude toward the undesirable proposition is simply an avowal or avowed belief, meaning a disposition to verbally affirm some content (Audi 1982). Some that it's an acceptance, and that belief doesn't entail acceptance (Cohen 1992). And some that it's a form of pretense, meaning imaginative pretense in the sense of make-belief or imagining (Gendler 2007).[4]

1. The Deflationary Strategy

Deflationists like Mele,[5] on the other hand, have attempted to bypass the static puzzle completely. By understanding self-deception as simply the product of biased information processing, they argue that we aren't required to attribute neither contradictory attitudes to the self-deceived, nor a tacit recognition encoded in terms of unconscious belief, but only a motivationally biased, false belief in the desirable proposition. The undesirable proposition, which motivates the self-deception, by their accounts, isn't believed by the self-deceived. This characterization can be seen in Mele's jointly sufficient conditions for entering self-deception in acquiring a belief that p:

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.
4) The body of data possessed by S at the time provides greater warrant for not-p than for p. (2001, p. 51)

[4] Again, I will not concern myself here with the problems raised by the attribution of non-belief cognitive attitudes toward the desired proposition in the context of self-deception. But see Van Leeuwen (2007) for criticism of Audi (1982).

[5] All further mentions of the deflationary view refer to Mele 2001, except when otherwise noted.

Noting that Mele's quartet of jointly sufficient conditions doesn't attribute to the subject any attitude toward the undesirable proposition (not-p), many philosophers have argued that they cannot account for the instability of self-deception. Robert Audi characterizes this instability as a kind of epistemic tension «ordinarily represented [...] by an avowal of p [...] coexisting with knowledge or at least true belief that not-p» (1997, p. 104). Closer to Mele's deflationism, Michael Losonsky argues that the self-deceived have the unwarranted, false belief that p, lack the true belief that not-p, and possess evidence for not-p that is active in their cognitive architecture. Importantly, his characterization includes the attribution of «some kind of recognition of the fact that the available evidence warrants the undesirable proposition more than the desirable one» (1997, p. 122, my emphasis). In this way, Losonsky means to supplement Mele's conditions in order to account for the conflict manifested, for instance, in recurrent or nagging doubts. Similarly to Audi, Mike W. Martin argues that «although self-deception does not involve fully conscious contradictory beliefs, typically it does involve a cognitive conflict, for example, suspecting p and believing not-p» (1997, p. 123). Likewise, Kent Bach contends that in self-deception, «unlike blindness or denial, the truth is dangerously close at hand» (1997, p. 105). Moreover, he observes that self-deception «ordinarily involves more than a one-shot mistreatment of the evidence. It involves repeated avoidance of the truth» (1997, p. 105). Finally, Eric Funkhouser has pressed on the point that the presence of avoidance behavior that points against the avowed belief is conceptually required for self-deception, noting that the self-deceived «engage in behavior which reveals that they know, or at least believe, the truth (not-p)» (2005, p. 303).

Mele responds to his critics by calling attention first to the fact that his jointly sufficient conditions don't entail that there is no tension in self-deception, and second to the fact that he hasn't anywhere claimed that self-deception normally is tension-free. He further contends that satisfying his four conditions may often involve psychic tension. He means to disarm his critics by pointing out that tension isn't conceptually necessary for entering self-deception in acquiring a belief that p (for which he doesn't offer a separate argument). Let's assume, for the sake of the argument, Mele's postulate that tension really isn't conceptually required, provided it's understood that it's a feature of many (if not most) cases. The question that needs to be answered is how could Mele account for inconsistent behavior?

He starts sketching an answer to this question when responding to critics, such as Losonsky, that maintain that his fourth condition is too weak and argue for a strengthened version that attributes to S a recognition that the body of data possessed by S at the time provides greater warrant for not-p than for p. The main reason for this contention is that, without such recognition, the self-deceived would have no reason to treat data in a biased way, since the evidence available would not be viewed as a threat in the first place, and consequently the self-deceived would not engage in motivationally biased cognition (avoidance behavior being one of the ways in which this is manifested). Mele notes that some theorists such as Donald Davidson (1985), and Harold Sackeim and Ruben Gur (1997) have concluded from this that when one is self-deceived in believing that p, one must be aware that one's evidence favors not-p. Mele's response has precisely this *awareness* in mind rather than simple *recognition*. He rightly points out that postulating such awareness places excessive demands on the self-deceived, since

> motivation can prime and sustain the functioning of mechanisms for the cold biasing of data in us without our being aware, or believing, that our evidence favors a certain proposition. Desire-influenced biasing may result both in our not being aware that our evidence favors not-p over p and in our acquiring the belief that p. [...] In each case, the person's evidence may favor the undesirable proposition; but there is no need to suppose the person is aware of this in order to explain the person's biased cognition. (2001, p. 53)

First, Mele's contention that motivation (i.e., our desire that p) can prime and sustain the functioning of unmotivated biasing mechanisms (some of which could be the availability heuristic and the confirmation bias) is plausible but misdirected.[6] The behavior we wish to explain by appeal to some sort of recognition on the part of the self-deceived is that expressed through the

[6] The availability heuristic refers to the tendency manifested when we form beliefs about the frequency, likelihood, or causes of an event, namely, that we «often may be influenced by the relative availability of the objects or events, that is, their accessibility in the processes of perception, memory, or construction from imagination» (Nisbett & Ross 1980, p. 18). The confirmation bias refers to a tendency manifested when we test a hypothesis, namely, that we tend to search more often for confirming than for disconfirming instances and to favor information that confirms our hypotheses regardless of whether the information is true (1980, pp. 181-82). For more detail on the different "cold" or unmotivated biasing strategies used by the self-deceived, see Mele 2001, pp. 28-9.

manifestation of motivated biasing mechanisms, especially selective focusing and attending, and selective evidence-gathering.[7]

Second, it isn't clear how motivation alone could function as Mele wants it to. Suppose I have a desire that this paper be accepted for publication. Would this suffice for me to avoid evidence that it won't? No. In order for that to happen, I would need a desire that this paper be accepted, coupled with a cognitive representation (let's leave it at that for the time being) that it won't or at least might not be accepted. In this way I would be motivated to avoid evidence that it won't (through the techniques mentioned) in order to avoid the distress involved in recognizing the evidence's weight. This doesn't necessarily imply the attribution of awareness, since, as Mele (2001, p. 80) himself has proposed, the priming of the biasing mechanisms could occur in a subpersonal level. Jeffrey Foss (1997) makes the similar point that conative attitudes like desire have no explanatory force without associated cognitive attitudes like beliefs. Mele (2001, p. 23) sees Foss' claim that motivational states must be linked to information states to explain behavior as an overgeneralization from a theory of intentional action, and points out that empirical evidence (e.g., Kunda 1990) proves that desires can generate behavior without being backed or accompanied by cognitive attitudes. Let's put aside the merit of Mele's answer to Foss, and assume that the biased treatment of evidence by the self-deceived isn't a product of an intentional project. The question raised by Mele's approach still remains unanswered: how can a desire that p, unaccompanied by some sort of recognition that not-p, lead one to avoid contact with the evidence that not-p?

Suppose we downgrade *recognition* to *information* (encoded in the mind of the self-deceived). There is evidence in the external world that indicates that not-p is true. To reiterate: unless some of this evidence that corroborates not-p

[7] Selective focusing/attending refers to the fact that our «desiring that p may lead us both to fail to focus attention on evidence that counts against p and to focus instead on evidence suggestive of p» (Mele, 2001, p. 26), and this behavior may or may not be intentional. Selective evidence-gathering refers to the fact that our «desiring that p may lead us both to overlook easily obtainable evidence for not-p and to find evidence for p that is much less accessible» (Mele, 2001, p. 27). This may be analyzed as «a combination of hypersensitivity to evidence (and sources of evidence) for the desired state of affairs and blindness [...] to contrary evidence (and sources thereof)» (Mele, 2001, p. 27). For more detail on the different "hot" or motivated biasing strategies used by the self-deceived, see Mele 2001, pp. 26–7). Literature on "selective exposure" is reviewed in Frey 1986.

is or might be true is encoded in the mind of the self-deceived (however inaccessible to consciousness, and however it's encoded), the self-deceived would have no motivation to avoid the evidence in the first place. A phenomenon so described would be but a case of wishful belief or wishful thinking. The resistance to the evidence present in self-deception would remain unaccounted for. Foss' use of "information states" encoded in the mind is helpful: he postulates neither belief nor intention. Neither do I. What I propose is that some information has to be encoded in the mind of the self-deceived.

While awareness seems to require that the subject has conscious access (most likely encoded as belief) that the evidence in his possession favors the undesired proposition, recognition could be interpreted as a subconscious, subdoxastic state, such as a mere suspicion that not-p is true. While it might seem that Mele supposes that, although the evidence the self-deceived possess favors the undesirable proposition, it isn't in anyway encoded in their mind, he does provide a complementary attribution to account for the conflicted behavior of the self-deceived. This is manifest in his analysis of Amélie Rorty's famous illustration of self-deception.

> Dr. Androvna, a cancer specialist, has begun to misdescribe and ignore symptoms of hers that the most junior premedical student would recognize as the unmistakable symptoms of the late stages of a currently incurable form of cancer. She had been neither a particularly private person nor a financial planner, but now she deflects her friends' attempts to discuss her condition and though young and by no means affluent, she is drawing up a detailed will. Although she has never been a serious correspondent and reticent about matters of affection, she has taken to writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon. (1988, p. 11)

Rorty's description of the case includes the safe assumption that Androvna lacks the conscious belief that she has cancer, so her behavior seems to require that she believes "deep down" that she is ill, or in Mele's (2001, p. 72) terms, that she has a type 1 cancer belief. Mele's answer to this, on the other hand, is that she consciously believes that there is a significant chance that she has cancer without also believing that she has it, i.e., she has a type 2 cancer belief. Hence, Mele's solution to the problem of accounting for the conflicted behavior of the self-deceived can be analyzed as the conjunction of the following attributions:

1) S believes that p.
2) S believes that there is a significant chance that not-p.

Mele rightly notes that we have and act on many type 2 beliefs, while also calling attention to the fact that there isn't comparably weighty evidence of type 1 beliefs. However, as Paul Noordhof (2009) points out, it's utterly unclear that Androvna's conscious belief that she doesn't have cancer would survive for long against a belief that there is a significant chance that she does have cancer. Nevertheless, let's assume, for the sake of the argument, that there may be circumstances in which both beliefs may be held simultaneously. The first crucial question concerning Mele's solution is whether the attribution of type 2 beliefs to the self-deceived can account for the inconsistencies in their behavior. An answer to this question will depend on working out the remaining details of Mele's solution.

Further steps in Mele's answer to the problem presented by the conflicted behavior of the self-deceived is found in his response to Tim Dalgleish's suggestion that «an individual can hold a propositional belief p while simultaneously having a higher-order emotional understanding of the situation consistent with not-p» (1997, p. 110). That is to say, someone might believe that p while also having a sense that not-p. Mele replies by asking whether this "sense" amounts to or encompasses a belief or «merely a suspicion that [not-p] or a belief that there is evidence that [not-p]» (2001, p. 79). Moreover, he argues that the conflicted behavior of the self-deceived «can be accounted for on the alternative hypothesis that, while believing that [p][...] self-deceivers also believe that there is a significant chance they are wrong about this» (2001, p. 80). We have just seen Mele appeal to this complementary belief, but this time he adds: «The mere suspicion that [not-p] does not amount to a belief that [not-p]. And one may entertain suspicions that p while believing that not-p» (2001, p. 80). Hence, we are entitled to add a new alternative to attribution number 2, namely:

3) S suspects that not-p.

I assume that Mele would be satisfied in attributing a conjunction of either 1 and 2 or 1 and 3 to the self-deceived. The second crucial question concerning Mele's solution is whether attributions 2 and 3 are different kinds of

attribution, or just different wordings of the same attribution. While we can safely assume that "suspicion" is understood by Mele to be a cognitive attitude, should we understand it as a kind of attitude on its own?[8] While Mele speaks of "a suspicion that [not-p] or a belief that there is evidence that [not-p]," it isn't absolutely clear whether or not he is equating these two attitudes. However, supposing that suspicion is to be taken as a kind of cognitive attitude on its own (and presumably a subdoxastic attitude), this would consist in falling back on one of the kinds of explanation eschewed by deflationists, namely, the strategy of postulating different kinds of attitudes toward p and toward not-p to account for conflicted behavior without attributing contradictory beliefs. What is more important, this would betray a tacit commitment to the idea that self-deception can't be made sense of without somehow attributing some kind of recognition (however it's encoded) to the self-deceived in order to account for the inconsistencies in their behavior.[9] Where Pears postulates different subagencies and Davidson different compartments to hold contradictory beliefs, and where Audi and others postulated an unconscious belief coupled with a subdoxastic attitude, Mele would postulate a subdoxastic attitude coupled with a conscious belief. This solution would not really be as parsimonious as deflationary theories want to be. Mele and others would have to supplement such an account by making explicit what kind of attitude they are referring to by "suspicion" (or whatever), why it should not be understood in doxastic terms, and, perhaps most importantly, how that subdoxastic attitude would be able to override belief and (at least sometimes) generate behavior.[10]

[8] Merricks (2009) is the only philosopher I know who raises the specific question of whether suspicion is a propositional attitude. His view is that if someone's attitude has a truth-value, then that attitude is a propositional attitude. So the suspicion Mele attributes to the deeply conflicted self-deceived would be, in Merricks' view, a propositional attitude. This much seems very plausible and uncontroversial. But Merricks doesn't investigate the further question of whether suspicion is its own kind of propositional attitude, or whether it is reducible to belief.

[9] Mele's is only a case in point. Other deflationist theorists try to sketch similar solutions and also end up attributing either a «recognition of the evidence as more or less establishing the contrary [of p]» (Johnston, 1988, p. 75), or a «suspicion» (Van Leeuwen, 2007, p. 428, fn. 19) or «uncertainty» (Barnes, 1997, pp. 42-3) on the part of the self-deceived toward the undesirable proposition. My criticism of Mele's position can be extended to these other deflationist theorists as well.

[10] A question Van Leeuwen (2007) raises concerning the attribution of avowal with respect to the self-deceptive belief. The attribution of suspicion with respect to the undesirable proposition raises a

While other deflationists may understand suspicion to be its own kind of cognitive attitude, Mele gives us reason to suppose that his use equates suspicion and belief with a degree of confidence short of certainty when he speaks of "a belief that there is evidence that [not-p]." In contrast to the mysterious use of "suspicion" as a kind of cognitive attitude and the questions this raises for the very intelligibility of such an explanation, probabilistic approaches to belief are at least much more straightforward. The self-deceived on Mele's account would harbor a belief that p and, at the same time, would (at least in some cases) harbor a belief that there is a chance that not-p. Remember that this was the way he worked out Rorty's Androvna example, and Noordhof's point that the belief with the lower degree of confidence (not-p) would plausibly not survive given the simultaneous presence of the belief with the higher degree of confidence (p). The problem now is, it isn't even clear in what exactly this mental state would consist. The third crucial question concerning Mele's solution is whether he is talking about two distinct beliefs, or rather about one belief with a degree of confidence between 0.5 and 1. If a person believes that p but at the same time isn't quite sure or "suspects" otherwise (i.e., believes that there is evidence that not-p), should we attribute to her a pair of contradictory beliefs (albeit with different degrees of confidence) or just one belief that p with a degree of confidence below 1?

The first of these options, namely, simultaneously making the attributions 1 and 2, engenders its own version of the static puzzle of self-deception: how can someone hold a belief that p and a belief that not-p (albeit of different degrees of confidence) at the same time? One of the key explanatory burdens of which Mele wishes to relieve his account of self-deception would be resuscitated, and the only way out of this would be to postulate at least a mild functional division along the lines proposed by Davidson. I will take it as an exercise in interpretive charity that Mele doesn't want to fall back on the division strategies he forcefully criticizes. I propose that the best way to understand his appeal to suspicion is as a diminishing of the confidence of the self-deceived in their self-deceptive belief that p. The conflicted behavior of the self-deceived would, on this account, be explained by the wavering of their confidence in the

similar question, since, however the undesirable proposition is encoded, endorsement of it is variously manifested in behavior.

belief that p. Mele's solution, then, would be characterized by the following attribution:

4) S believes that p with a degree of confidence that alternates between 0.5 and 1.

This shifting back and forth could easily be explained as the product of the subject's relationship with the threatening data (e.g., through the activation of certain memories, through the admonishing of relatives and friends, through direct contact with the evidence, etc.). One's confidence in the self-deceptive belief would fluctuate and thus manifest itself in behavior that at one time would point toward a higher, and at other times toward a lower, confidence in p. However, because Mele attributes only a suspicion that not-p to the self-deceived, he would still be hard-pressed to explain behavior that points toward a high degree of confidence in not-p, which would indicate that the degree of confidence in p sometimes drops below 0.5. Take Androvna's case, for example. Her confidence in not-p (i.e., that she has cancer) is apparently higher than her confidence in p when she is «drawing up a detailed will,» or «writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon» (Rorty 1988, p. 11). The fourth crucial question concerning Mele's solution, then, is whether we can really account for the inconsistencies in the behavior of the self-deceived by attributing to them a belief with a degree of confidence which wavers between 0.5 and 1. The answer is no. The only way to account for the relevant behavior by attributing to the self-deceived only one belief would be by making the following attribution:

5) S believes that p with a degree of confidence that alternates between 0 and 1.

What this means is that the theorist that opts for a single, determinate belief attribution will depart from the deflationist's original intent of attributing the belief that p and will, if she aims at accounting for inconsistent behavior, attribute an intermittent belief to the self-deceived. How could we make sense of this? Supposedly, when the subject manifests p-behavior, we would attribute to her the belief that p. When she manifests not-p-behavior, we would attribute to her the belief that not-p. This doesn't, of course, imply that the subject holds the belief that p and the belief that not-p simultaneously. However, this is, I suggest, a complete breakdown of the ordinary way of understanding and

practicing belief attribution. Such an attribution has absolutely no explanatory or predictive power and makes out "belief" to be purposeless.

Setting out from the failure of deflationary accounts of self-deception in making precise attributions to the self-deceived, I want to claim that we should recognize in self-deception one of the innumerable examples that corroborate the maxim that «ordinary language breaks down in extraordinary cases» (Austin, 1979, p. 68).

2. Toward a Dispositionalist Approach

Eric Schwitzgebel (2001) has rightly recognized that there are countless cases in which a simple yes or no answer to the question "Does S believe that p?" doesn't seem to be available, and that they can have a wide variety of causes. Self-deception is just one of these, among others such as implicit associations (Schwitzgebel, 2010) and delusions (Schwitzgebel forthcoming). From the presence of these cases, Schwitzgebel draws the following conclusion:

> For any proposition p, it may sometimes occur that a person is not quite accurately describable as believing that p, nor quite accurately describable as failing to believe that p. Such a person, I will say, is in an "in-between state of belief" (Schwitzgebel, 2001, p. 76).

The reason such a person isn't accurately describable is that she doesn't accurately fit the stereotype for believing that p, while at the same time also failing to accurately fit the stereotypes for other intentional attitudes, such as the stereotypes for believing that not-p, imagining that p, etc. It must be noted, however, that the label "in-between belief" doesn't pick out a particular kind of state that someone is determinately in. As Maura Tumulty notes, "in-between belief" is only meant as «a convenient way of referring to the fact that a particular subject fails fully to meet any relevant folk-psychological stereotype» (forthcoming). The widespread presence of problematic circumstances for belief attribution such as those of self-deception encourages the development of an account of belief that allows us to talk intelligibly about such in-between states – that allows us to say more than just that the subject "sort of" believes something. An approach of this kind is already surfacing in the literature on

delusions,[11] but it has been almost completely neglected in the literature on self-deception.[12] I contend that the explanatory failure of all the accounts of self-deception that have been proposed so far hinge precisely on unrealistic assumptions about the limits of folk psychology. With Funkhouser, I want to claim that our failure in trying to characterize precisely the mental state of the self-deceived «is not due to our limited epistemic perspective; rather, it is a real indeterminacy» (2009, p. 9). That is to say: a real indeterminacy in our folk-psychological concepts. And with Foss, I want to point out that since beliefs and desires «cannot be independently observed somewhere in the head [...] the only constraint on their attribution is the cogency of the resulting explanation itself» (1997, p. 112).

Of course, in accounting for self-deception and other in-between states we should strive to complement the negative attitude toward the attribution of belief in complex cases that I am recommending with a positive methodology for the best possible description of these cases. With Schwitzgebel, I think our best bet is to develop explanations of these phenomena that set off from an

[11] Bayne and Pacherie (2005) first sketched an account of delusional belief inspired by Schwitzgebel (2002). For criticism of their account, see Tumulty 2011. For the idea that delusional states should be included in the category of in-betweenish states, see Schwitzgebel (forthcoming) and Tumulty (forthcoming). For the related idea that there is no fact of the matter concerning what delusional subjects believe, see Hamilton 2006.

[12] Schwitzgebel was perhaps the very first one to point this out: «In the self-deception literature the option of refusing to say that either "yes the self-deceived person believes the unpleasant proposition" or "no she doesn't" is surprisingly uncommon. One sees this view, perhaps, in H. O. Mounce's (1971) paper on the subject, and Mele describes it as an option in a review article on self-deception (1987b), although he neither accepts the idea nor specifically addresses it in his positive work on the topic» (1997, p. 306). The only reference to the approach I am suggesting that I could find within the self-deception literature is in Bayne and Fernández (2008, p. 8): «A second response to the problem takes issue with the assumption that it is not possible for an agent to believe p and believe not-p at one and the same time. According to some approaches to belief, it is possible for an agent to have inconsistent beliefs at one and the same time, as long as the beliefs in question have different triggering conditions (Lewis 1986, Schwitzgebel 2002). The dispositions distinctive of believing p will be activated by one triggering condition, whilst those distinctive of believing not-p will be activated by other triggering conditions.» But it's important to note that Bayne and Fernández actually misconstrue Schwitzgebel's account, as they read him as proposing a model for how to account for contradictory (or rather, conflicting) beliefs, where it actually proposes a model for how to account for conflicting dispositions. In hard cases, the attribution of conflicting beliefs is substituted by an attribution of the dispositions manifested, because in hard cases some of these dispositions point toward different directions (p, not-p) and can't be made sense of by an attribution of a determinate belief state. See Schwitzgebel 2010, p. 544.

account that identifies believing with being disposed to act and react in various ways in various circumstances. Better yet: an account which is built upon a broad dispositional base. Schwitzgebel suggests that one way of articulating this is to say that «beliefs are not "single track" dispositions but rather multi-track» (2010, p. 533) – or in Gilbert Ryle's terminology, that they «signify abilities, tendencies or pronenesses to do, not things of one unique kind, but things of lots of different kinds» (1949, p. 118). What should we say, then, when a person appears to only partly possess the relevant dispositional structure? Here is what I take to be the core of Schwitzgebel's answer:

> If to believe is to possess a multi-track disposition or a broad-track disposition or (as I myself prefer to put it) a cluster of dispositions (which can include cognitive and phenomenal dispositions as well as behavioral ones), then there will be in-betweenish cases in which the relevant disposition or dispositions are only partly possessed. And if we treat such cases analogously to other cases of the partial possession of multi-track or broad-track dispositional structures, then we should say of such cases that it's not quite right, as a general matter, either to ascribe or to deny belief simpliciter – though (as in the other examples) certain limited conversational contexts may permit simple ascription or denial. Belief language starts to break down; the simplifications and assumptions inherent in it aren't entirely met; in characterizing the person's dispositional structure we may have to settle for lower levels of generality. (2010, p. 535)

After descending to a lower level of description than that of "believes that p," and articulating the subject's dispositional structure in the finest possible detail we can, we may complement our description by matching certain dispositional patterns with certain belief stereotypes, or by investigating the etiology of the relevant phenomenon to propose an answer as to why and how the mixed set of dispositions is acquired, etc. But having done that, we will have done what is possible for us to do (at least for now). An account of self-deception developed strictly along these theoretical assumptions has not yet appeared, but all accounts that have been developed provide us with a vast array of useful data concerning the dispositional structure of subjects engaged in self-deception. All we need to do now is to stop worrying about what the self-deceived really believe, and focus on refining our descriptions of the dispositional make-up of the self-deceived.

3. Brief Conclusion

In many everyday cases it is clear what a person believes. On the other hand, we have seen that in self-deception it is not at all clear if the subject believes that p, believes that not-p, suspects that not-p, etc. Trying to make sense of the most parsimonious accounts of self-deception leads us to the same problems that the traditional accounts have raised. No account so far has been able to make sense of the inconsistency and instability suggested by the behavior of the self-deceived, which is precisely one of the reasons why self-deception is so interesting. It helps us notice the limits of application of our folk-psychological concepts, and pushes us to come up with more refined ways to analyze our psychological attitudes toward propositions. However, while the correct response is to refrain from either attributing or denying belief when the dispositions the subject manifests do not warrant a determinate attribution, we are able to come up with explanatory and predictive descriptions of the behavior and dispositional structure of the self-deceived, and this is what we should be doing.

REFERENCES

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

Audi, R. (1997). Self-Deception vs. Self-Caused Deception: A Comment on Professor Mele. *Behavioral and Brain Sciences*, *20*(1), 104.

Austin, J.L. (1979). The Meaning of a Word. In J.O. Urmson & G.J. Warnock (Eds.), *Philosophical Papers*. New York: Oxford University Press.

Bach, K. (1997). Thinking and Believing in Self-Deception. *Behavioral and Brain Sciences*, *20*(1), 105.

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

Bayne, T., & Pacherie, E. (2005). In Defence of the Doxastic Conception of Delusions. *Mind and Language*, *20*(2), 163–188.

Bayne, T., & Fernández, J. (2009). Delusion and Self-Deception: Mapping the Terrain. In T. Bayne & J. Fernández (Eds.), *Delusion and Self-*

Deception: Affective and Motivational Influences on Belief Formation. New York: Psychology Press, 1–21.

Cohen, L.J. (1992). *An Essay on Belief and Acceptance*. New York: Clarendon Press.

Dalgleish, T. (1997). Once More with Feeling: The Role of Emotion in Self-Deception. *Behavioral and Brain Sciences*, *20* (1), 110–111.

Davidson, D. (1982). Two 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: Blackwell, 138–148.

Foss, J.E. (1997). How Many Beliefs Can Dance in the Head of the Self-Deceived? *Behavioral and Brain Sciences*, *20* (1), 111–112.

Frey, D. (1986). Recent Research on Selective Exposure to Information. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology*. New York: Academic Press, 41–80.

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

Funkhouser, E. (2009). Self-Deception and the Limits of Folk Psychology. *Social Theory and Practice*, *35* (1), 1–13.

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

Hamilton, A. (2006). Against the belief model of delusion. In M.C. Chung, K.W.M. Fulford, & G. Graham (Eds.), *Reconceiving Schizophrenia*. Oxford: Oxford University Press, 217–234.

Heil, J. (1993). Going to Pieces. In G. Graham & L. Stephens (Eds.), *Philosophical Psychopathology: A Book of Readings*. Cambridge, MA: MIT Press, 111–133.

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.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Lewis, D.K. (1986). *On the Plurality of Worlds*. Oxford: Blackwell.

Losonsky, M. (1997). Self-Deceivers' Intentions and Possessions. *Behavioral and Brain Sciences*, *20*(1), 21–122.

Martin, M.W. (1997). Self-Deceiving Intentions. *Behavioral and Brain Sciences*, *20* (1), 122–123.

Mele, A.R. (1987a). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. New York: Oxford University Press.

Mele, A.R. (1987b). Recent Work on Self-Deception. *American Philosophical Quarterly*, *24*(1), 1–17.

Mele, A.R. (1997a). Real Self-Deception. *Behavioral and Brain Sciences*, 20 (1), 91–102.

Mele, A.R. (1997b). Understanding and Explaining Real Self-Deception. *Behavioral and Brain Sciences*, *20*(1), 127–134.

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

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

Merricks, T. (2009). Propositional Attitudes? *Proceedings of the Aristotelian Society*, *109*, 207–232.

Mounce, H.O. (1971). Self-Deception. *Proceedings of the Aristotelian Society*, *45*, 61–72.

Nisbett, R., & Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice-Hall.

Noordhof, P. (2009). The Essential Instability of Self-Deception. *Social Theory and Practice*, *35* (1), 45–71.

Pears, D. (1984). *Motivated Irrationality*. Oxford: Clarendon Press.

Price, H.H. (1969). *Belief: The Gifford Lectures Delivered at the University of Aberdeen in 1960*. Muirhead Library: George Allen and Unwin.

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.

Ryle, G. (1949). *The Concept of Mind*. New York: Barnes & Noble.

Sackeim, H.A., & Gur, R.C. (1997). Flavors of Self-Deception: Ontology and Epidemiology. *Behavioral and Brain Sciences*, *20*(1), 125–126.

Schwitzgebel, E. (1997). *Words about young minds: The concepts of theory, representation, and belief in philosophy and developmental psychology*. University of California, PhD dissertation.

Schwitzgebel, E. (2001). In-between Believing. *Philosophical Quarterly*,*51*(202), 76–82.

Schwitzgebel, E. (2002). A Phenomenal, Dispositional Account of Belief. *Noûs*, *36*(2), 249–275.

Schwitzgebel, E. (2010). Acting Contrary to Our Professed Beliefs, or the Gulf Between Occurrent Judgment and Dispositional Belief. *Pacific Philosophical Quarterly*, *91*(4), 531–553.

Schwitzgebel, E. (forthcoming). Mad Belief? *Neuroethics*.

Tumulty, M. (forthcoming). Delusions and Not-Quite-Beliefs. *Neuroethics*.

Tumulty, M. (2011). Delusions and Dispositionalism about Belief. *Mind and Language*, *26*(5), 596–628.

Van Leeuwen, D.S.N. (2007). The Product of Self-Deception. *Erkenntnis*, *67*(3), 419–437.